Probabilistic modelling for assessment of exposure via drinking water

Final Report of Project Defra WT1263 / DWI 70/2/273

Cranfield UNIVERSITY

Centre for Environmental
Risks and Futures
November 2012

Probabilistic modelling for assessment of exposure via drinking water

Final Report of Project Defra WT1263 / DWI 70/2/273

Prepared for the Drinking Water Inspectorate by
Dr David Parsons, Centre for Environmental Risks and Futures, Cranfield University
Dr Mick Whelan, Centre for Environmental Risks and Futures, Cranfield University
Dr Ruth Bevan, Institute of Environment and Health, Cranfield University

Centre for Environmental Risks and Futures
Building 42
Cranfield University
Cranfield
Bedford
MK43 9PB
www.cranfield.ac.uk

Probabilistic modelling for assessment of exposure via drinking water

Final Report of Project Defra WT1263 / DWI 70/2/273

David Parsons, Mick Whelan, Ruth Bevan
Cranfield University

Units of intake

Throughout this report we follow the convention used in the tap water consumption surveys (Accent, 2008; MEL Research, 1996; Hopkin & Ellis, 1980) and the majority of papers by expressing intakes in l/d, meaning litres per person per day. Similarly, we follow the National Diet and Nutrition Survey (Bates *et al.*, 2011) and the usual convention for reference nutrient intakes (Buttriss, 2000) by expressing the intake of specific chemicals in mg/d (or µg/d) meaning milligrams (or micrograms) per person per day, unless they are specifically quoted on a unit body mass basis, for example Acceptable Daily Intakes in mg $kg^{-1}d^{-1}$ (or µg $kg^{-1}d^{-1}$).

Executive summary

Extensive data are available on the concentrations of a wide range of chemicals in drinking water and also on the amount of water consumed by the public. The approach used to set drinking water quality standards and to do risk assessments only uses part of this data. Typically the highest measured concentration is combined with a high end consumption value to give a deterministic estimate of exposure. This can then be compared with a toxicological based intake that is considered to pose no significant risk over a lifetime's exposure. Such an approach affords a high degree of protection to consumers. However, it may be useful to know what the statistical distribution of exposure is likely to be, in order to understand the range of exposure within the population and to make better estimates of extreme exposure percentiles. Probabilistic assessments use the whole distribution for each variable to derive a distribution for the intake of the chemical from which summary statistics may be derived, if required.

The main objective of this project was to develop and explore the potential benefits of probabilistic approaches to exposure assessment to chemicals by ingestion of tap water in England and Wales.

Several approaches to probabilistic modelling were reviewed to assess their suitability for this purpose: analytical methods, Monte-Carlo simulation, Bayesian networks and Markov-Chain Monte-Carlo methods. The best method to provide the results required from the available data was assessed to be Monte-Carlo simulation. The simulations were implemented using the statistical programming language R, which was the package used for the data analysis. Methods of performing the simulations in Excel™ were also demonstrated.

Data on intakes of tap water by the population was available from drinking water consumption surveys (DWCS) in 1978, 1995 and 2008 (referred to here as DWCS1978 etc), a survey for the Food Standards Agency (FSA) in 2001 (NDNS 2001) and a survey of consumption by children under 16 for the DWI in 2011 (DW16 2011). The coverage and level of detail available varied between these. The DWCS 1978 and DWCS 1995 surveys included adults and children, whereas DWCS 2008 and NDNS 2001 included only adults. The only results available from the DWCS surveys were in the form the frequencies of different intakes. The results for individual intakes were available from NDNS 2001 and DW16 2011, which gave other information, such as the age and weight of the individuals. The reports of the DWCS analysed the effects of these factors, but did not give the data. All the surveys used 7-day diaries to record intake, but with two different methods of assessing the volume consumed. DWCS 1978 and NDNS 2001 used measured or weighed amounts, whereas the others relied on standard vessel volumes and estimates of the proportion filled and drunk.

The mean tap water intake by adults appeared to increase from 1.084 l/d in DWCS 1978 to 1.275 l/d in DWCS 1995 and to 1.294 l/d in DWCS 2008. The first increase was statistically significant, but the second was not. The mean intake in NDNS 2001 was 1.103 l/d, which was significantly different from the results for DWCS 1995 and 2008. It should be noted that the differences in intake coincide with the different methods used

to estimate the volumes used by the surveys, so it is possible that the apparent change in intake is an artefact of the method of estimation of consumption. The mean intake by all children in DW16 2011 was 0.551 l/d

Intake by adults and children tended to increase with both age and body weight, but the variability within each weight or age group was very high. The results for males and females in different surveys were inconsistent: in DWCS 1978, NDNS 2001 and DW16 2011, males drank more tap water than females; in the other surveys they drank less. It is interesting to note that the difference in the results for adults coincides with the difference in the recording method.

Probability distributions were fitted to the water intake data for each survey using maximum likelihood estimation. Four distributions were tested: (1) lognormal, (2) normal fitted to the square root of intake (as a continuous analogue of the Poisson distribution), here referred to as *normal (square root)*, (3) Weibull and (4) gamma. None gave a consistently better fit than the others; the gamma distribution generally gave good results and was selected for the simulation. Distributions were fitted to the complete populations and to the groups by sex and age or weight where they were available.

Data on the concentrations of chemicals in drinking water was available from compliance data collected by the water companies and supplied to the DWI. Data sets containing iron, lead, selenium, sodium and manganese concentrations from 2004 and 2010 were used in the study. Additional data sets on lead from some of the intervening years and less detailed data on lead from 1994 were used to look at the trends in exposure to lead over time. Some data on copper from 2010 was also included. Several distributions were fitted to the data using maximum likelihood estimation. In most cases, the best fit was obtained by the lognormal distribution, with the exception of selenium, for which the exponential distribution fitted better.

The data sets presented two problems: most sets had very long tails containing a few high values and several had a large proportion of small values that were not precisely determined, but recorded as below the limit of detection (LoD), for which a value was given. The problem of data below the LoD, known as left-censored data, was dealt with by using a version of maximum likelihood estimation that was designed for censored data. The long tails of the distributions could not be fitted by any standard distributions: all of them underestimated the frequency of rare extreme values, which are potentially important when assessing exposure probabilistically. As the data sets were large (over 10,000 values), the method chosen for the simulation was to sample from the original data instead of using the fitted distributions for most of the range. The distributions were used to generate substitute values for those lying below the LoD. The simpler alternatives of substituting 0, LoD/2 or the LoD were also tested.

Reference values for intakes of the substances being considered were taken from authoritative UK and other sources. There were three types of reference values. (1) Reference Nutrient Intakes (RNIs) for required minerals giving the amount that is sufficient for about 97% of the population (or similar measures). These existed for all the substances other than lead. (2) Safe limits for long-term intake, such as Acceptable

Daily Intake (ADI) or Provisional Maximum Tolerable Daily Intake (PMTDI), for potentially harmful substances. These were found for all the substances except manganese. (3) Intakes from sources other than tap water, such as dietary intake.

The simulation was written in R with a bespoke user interface and run with every intake distribution (i.e. the distributions fitted to the full sample in each survey and, where possible, the groups by sex and age or weight) and the 2010 concentration data. Additional runs were carried out using some of the intake distributions with the earlier sets of concentration data to explore changes over time. In each iteration of the simulation, a pair of values was sampled randomly from the appropriate water intake distribution and concentration data set, as described above, and multiplied to give a value for chemical intake. Sampling from the two distributions was independent: that is, no correlation between water intake and concentration was considered. Each run consisted of 100,000 iterations of the simulation, from which summary statistics were derived, including the mean, median, 99th and 99.9th percentile.

The results of the simulations have shown that exposure to metals in tap water is highly variable. The 99.9th percentile exposure can be up to 45 times the mean and 200 times the median. It should be emphasised that the percentiles relate to the chance of individual daily exposures, not long-term intake.

The method of substitution for values less than the LoD had only moderate effects on the estimation of the mean and percentiles up to the 75th for those substances with a high proportion of samples reported at the LoD (e.g. lead and iron), and smaller effects on the statistics of the other substances. The higher percentiles were unaffected in all cases. For simpler exposure assessments, substitution by either the LoD or LoD/2 would probably give acceptable accuracy.

Exposure to iron, lead, selenium and manganese predicted by the simulations appear to have decreased by about 40% between 2004 and 2010 due to falling concentrations in tap water. For lead this is part of a long term trend, having previously decreased by 40% between 1994 and 2004. In contrast, the exposure to sodium appears to have increased slightly.

Comparing the predicted exposures with Reference Nutrient Intakes (for required nutrients) and Acceptable Daily Intakes or other recommended maximum intakes, we found for adults, using 2010 concentration data, that:

- For iron, selenium, sodium and manganese, the 99.9th percentile exposures were much less than the RNIs. In each case, the RNI is much lower than the ADI or similar upper limit.
- For copper, the 99.9th percentile exposure slightly exceeded the RNI and the intake from other sources, but the mean was very much smaller than the RNI, so tap water may occasionally make a significant contribution to the requirement for copper. The 99.9th percentile exposure was less than 10% of the PMTDI.
- For lead, the 99.9th percentile exposure was about 40% of the ADI in the worst case and the mean exposure was about 1% of the ADI. The ADI is being

superseded by BMDL values. The mean exposure was about 5% of the $BMDL_{10}$ for nephrotoxity, which lay between the 99th and 99.9th percentiles of exposure in the worst case. The ADI and the BMDL both relate to lifetime exposure, not acute effects, so the mean exposure is the most appropriate comparison.

For children under 16, using 2010 concentration data, we found that:

- For iron, the 99.9th percentile of predicted exposure was less than 5% of the RNI in all cases. In the worst case, the 99.9th percentile was less than 5% of the PMTDI.
- For selenium, the 99.9th percentile of predicted exposure was less than 20% of the RNI in most cases and less than 50% in the case of the youngest group. The mean exposure was less than 4% of the RNI. Thus, tap water may occasionally, but not persistently, be a significant contributor to nutrient intake. The 99.9th percentile is 0.01% of the Upper Safe Level in the worst case.
- For sodium, the 99.9th percentile of predicted exposure for the youngest group (worst case) was about 30% of the RNI and the mean was about 3% of the RNI. The intake from other sources normally exceeds the RNI. Therefore, tap water may occasionally, but not frequently, be a significant contributor to sodium intake.
- For manganese, the 99.9th percentile of predicted exposure was less than 2% of the adequate daily intake for all groups aged 4 years and upwards. For the 0–3 years group, the 99.9th percentile was less than 4% of the adequate daily intake for children aged over 6 months, but 7 times the adequate daily intake for babies up to 6 months. The data set is insufficient to allow this age group to be simulated separately, but it is possible that the required nutrient intake may occasionally be exceeded for babies up to 6 months. However, manganese has low acute toxicity, and no PMTDI has been set.
- For lead, the 99.9th percentile exposure was less than half of ADI (for lifetime exposure) for most groups, but exceeded it slightly for the lightest group. The mean exposure was about 3% of the ADI in the worst case and about 1% of the ADI for the other groups. The mean exposure was less than 6% of the $BMDL_{01}$ (for long-term exposure) for developmental neurotoxicity in most cases and less than 20% in the worst case. The $BMDL_{01}$ was between the predicted 99th and 99.9th percentiles for most groups and between the 95th and 99th percentiles in the worst case. Thus the probability of persistently exceeding this level is relatively small.

Similar methods could be applied to other substances found in tap water, such as Trihalomethanes (by-products of chlorination). There are significant routes of exposure other than ingestion for these chemicals, notably skin contact and inhalation when bathing. A more complex model would therefore need to be constructed to adequately represent exposure to these substances.

1. Introduction to probabilistic modelling of exposure

Extensive data are available on the concentrations of a wide range of chemicals in drinking water and also on the amount of water consumed by the public. The approach used to set drinking water quality standards and to do risk assessments only uses part of this data. Typically the highest measured concentration is combined with a high end consumption value to give a deterministic estimate of exposure. This can then be compared with a toxicological based intake that is considered to pose no significant risk over a lifetime's exposure. Such an approach affords a high degree of protection to consumers. However, it may be useful to know what the statistical distribution of exposure is likely to be, in order to understand the range of exposure within the population and to make better estimates of extreme exposure percentiles. Probabilistic assessments use the whole distribution for each variable to derive a distribution for the intake of the chemical from which summary statistics may be derived, if required.

The main objective of this project was to develop and explore the potential benefits of probabilistic approaches to exposure assessment to chemicals by ingestion of tap water in England and Wales.

The main sources of water intake data for adults are household surveys. For example, the Drinking Water Survey (Accent, 2008; Marsden, 2010), contains self-recorded data on the intakes of about 1500 adults from 1000 households in two periods in Spring and Summer 2008. The values are average daily intakes based on seven day diaries, presented as a frequency distribution (histogram), with the intake divided into 0.1 litre intervals. The available survey data are summarised in Section 2: most are only available as frequency distributions with summary statistics for demographic factors, but the individual (anonymous) records are available for some of the recent surveys.

Babies and children under 16 may be more sensitive than adults to some chemicals and their generally lower body weight may mean that their intake relative to their weight is higher than that of adults. During the project, the results of a survey of tap water intake by children less than 16 years old that was carried out for Defra/DWI in 2011–12 became available. (Ipsos MORI, 2012). Similar analyses were conducted using this data set.

Data on chemical concentrations come from the monitoring data supplied by water companies in England and Wales to the Drinking Water Inspectorate and consist of chemical concentrations in samples for from each company. These are considered in detail in Section 4. The number of samples depends on the chemical, ranging from 9,000 per year for mercury to 45,000 per year for iron, from the whole country, although for some substances up to 90% are below the limit of detection.

Similar problems have been encountered by other government departments and their approaches were reviewed by the Interdepartmental Group on Health Risks from Chemicals into reports. Their findings are summarised in section 1.2.

Unless otherwise stated, all analyses conducted within the project were carried out using the statistical programming language R (Crawley, 2007), which is freely available (CRAN, 2012).

1.1 Potential approaches to drinking water exposure assessment

1.1.1 Analytical method

Exposure to a chemical contaminant via drinking water is the product of water intake rate and concentration. It is often possible to describe the shape of the frequency distribution using a mathematical function called a probability density function (pdf). There are many different shapes of pdf representing a wide range in the relationship between a phenomenon and its probability. The integral of the pdf with respect to the variable (i.e. the cumulative probability of a variable occurring which is less than a given value) is called the cumulative distribution function (CDF), or simply the distribution function.

If the distributions of concentration and tap water intake could both be represented by parametric distributions, such as normal (Gaussian), lognormal or exponential distributions, it would be possible to derive the distribution of exposure from the *convolution product* (see Glossary) of the two distributions.

To illustrate the complexity of calculating the convolution product, consider first the case of a few discrete values of intake, e.g. 0.5, 1.0 and 2.0 l/d with probabilities p_5, p_{10} and p_{20}, and discrete values of concentration, e.g. 2, 4 and 8 mg/l with probabilities q_2, q_4 and q_8. It is easy to see that there are three combinations of intake and concentration that give an exposure of 4 mg/d with a combined probability of $p_5q_8+p_{10}q_4+p_{20}q_2$. Similarly, there are two combinations giving 8 mg/d and so on.

With many discrete values, there would be many possible combinations and summations of many terms. When discrete values are replaced by continuous distributions, these sums become infinite and are replaced by integrals. In general, these integrals cannot be solved analytically; for example, the product of two normal distributions is a *Bessel function of the second kind*, which has no closed form. It would be possible to use numerical integration, but there are few benefits over the methods discussed later to justify the complexity of this approach.

In practice, it was found that the data on the concentration of most chemicals could not be well described across its full range by standard pdfs, so there was no reason to pursue this approach.

Similar problems would apply to multiplying frequency distributions in the form of histograms. Again, begin with a simplified case in which intake is given by a histogram and concentration by discrete measurements. Suppose we consider two intervals for intake: 1.0–1.5 l/d and 1.5–2 l/d. Similarly consider two concentration measurements: 2 mg/l and 2.5 mg/l. Multiplying these we get four intervals for exposure: 2.0–3.0, 3.0–4.0, 2.5–3.75 and 3.75–5.0 mg/d. The intervals are of different lengths and overlap, so separating exposure into intervals and calculating the frequency in each becomes complex when the full distribution is included. Replacing the discrete measurements of concentration with a histogram for the distribution compounds the problem.

1.1.2 Monte-Carlo simulation

Monte-Carlo simulation (MCS) is probably the most widespread approach to stochastic (probabilistic) modelling. It is typically used where it is possible to construct a mechanistic model of a process or system, but some of the inputs are known to be subject to natural *variability* or where there is *uncertainty* about the true values of some of the inputs or parameters. The calculations are performed iteratively, typically thousands of times. In each iteration, the values of all the variable or uncertain inputs and parameters are drawn independently from appropriate distributions to produce distributions for the output variables.

For the case of exposure via drinking water, the model is very simple: the product of intake and concentration. A single iteration of the model takes one randomly selected value of intake and one randomly selected value of concentration and multiplies them to give an exposure. MCS would repeat this for a large number of iterations to create a virtual sample from the population.

It is important to note that although the selection of each input variable in each iteration is random: the probability of selecting variable values is governed by the shapes of the probability distributions employed, so some values are more likely to be selected than others. The selection of the input probability distributions is usually made by fitting pdfs to observed data.

Most of the available tools (see 1.1.3) include a wide range of standard distributions. Some care may be needed when using distributions that can take values over a very wide range to ensure that the values remain reasonable. For example, there is always a possibility of a value sampled from a normal distribution being negative.

It is also possible to use the original data as the source and simply draw samples from it at random. The disadvantage is that this will only ever produce values from the discrete set in the data, but it is a reasonable approach if the data set is large.

Another approach is to construct an empirical cumulative distribution function (ECDF) from the data. The data points are placed in ascending order and assumed to represent equal quantiles (see Glossary) from the distribution.

In the ECDF, the points are joined by a series of steps. It is then possible to treat this as a distribution function from which to sample. (Strictly speaking, it is the inverse of the CDF that is used.) However, when using a step function this is equivalent to the previous method and it will also only give values that were present in the original data. Some packages provide a function that will derive arbitrary quantiles from the data, effectively interpolating between the points.

If the data is available only as a histogram, choosing a representative from each class, such as the central value, would result in only a very limited number of values being used. An alternative is use uniform random samples from each interval, but this would introduce a bias, especially for a skewed distribution. It is usually preferable to fit a distribution to the histogram.

The advantages of MCS are that it can be used with any model, for any input distributions and that the model can be constructed using commonly available software, such as Microsoft Excel™ (see 1.1.3). The main disadvantage is the need for repeated, and possibly time consuming, simulations to derive the results, which will vary slightly each time the model is run. It is also difficult or impossible to include some features of probabilistic modelling, such statistical inference.

1.1.3 Tools for Monte-Carlo simulation

A popular approach to MCS is to build the model without uncertainty in a spreadsheet, such as Microsoft Excel™, then use third-party tools, such as @RISK™ (Palisade Corporation, 2007) and Crystal Ball™, to generate the input distributions, collect the output, display it graphically and compute summary statistics. Alternatively, there are free or open source alternatives for use with Excel, such as Simulación (Varela, 2011) and Monte-Carlito (Auer, 2012).

Simulations can also be constructed using specialised tools (e.g. GoldSim™), statistical programming languages (e.g. R), or written in general-purpose programming languages, such as Fortan, C++ or Java. Specialised tools provide comprehensive frameworks for simulation, simplifying development and providing rich graphics, but are generally expensive, which limits the options for distributing models. Statistical languages usually include the necessary tools, such as random number generators and graphics, but are less suited to non-specialists. Some are commercial (e.g. SPSS and Genstat) and others are free (R), making widespread distribution possible. Using general-purpose languages requires the most effort, because they are not tailored to the problem, but gives great flexibility. They are probably most suitable when other constraints, for example integration with other systems or delivery as a web application, limit the use of specialised tools. Many free and commercial options are available.

All of these options are simply tools to achieve the same result. Provided the random number generators are of good quality, the results should be the same, within the limits of variation arising from the use of random numbers. We have used @RISK, Simulación and Monte-Carlito successfully in previous projects, so a simple exposure simulation was constructed in Excel with each of them to compare the performance and results.

Both @RISK and, Simulación are add-on packages containing libraries of probability distributions, methods of selecting the input and output variables, and a user interface to control the operation. Although the inputs are normally specified using distributions, it is straightforward to sample from a data set, by generating a random integer between 1 and the length of the data and using it as an index in a look up function. Using a distribution to substitute for concentration values less than the limit of detection, as discussed later, would be a little more complicated, but still feasible. They are similar in their styles of operation, although @RISK permits more input variables and has more sophisticated graphical output. Conversely, it was easier to access the raw output for further processing in Simulación than @RISK, because it was stored in a worksheet.

Monte Carlito is simply a set of macros written in Visual Basic for Applications™ rather than an add-in. This means that it is limited to the small set of distribution functions provided by Excel and the standard graphics are poor. However, it can be fully integrated into a spreadsheet without the need for additional packages and the source code is accessible, so it can be tailored to the requirements of a particular application, provided that the terms of the GNU General Public Licence are adhered to.

A simple exposure simulation was constructed in Excel consisting of one cell for water intake drawn from a gamma distribution, one cell for contaminant concentration drawn from an exponential distribution and one cell to calculate the product of these two values, giving the exposure. The simulation was implemented using @RISK, Monte-Carlito and Simulación. With @RISK and Simulación the intake and contaminant distributions were defined by selecting the cells and using the plug-in menus to specify them as inputs with the appropriate distributions. The output was identified by selecting the cell and specifying it as an output using the plug-in menus In Monte-Carlito the inputs were created using standard Excel formulae that called the RAND() function. Its design required the input and output cells to be placed in a row with vacant space below it to receive the summary statistics. Each simulation was run for 65,000 iterations and timed. It required 6 seconds in @RISK, 30 seconds in Monte-Carlito and 70 seconds in Simulación. For comparison, the same model implemented in R ran in about 0.1 seconds (timed over 6,500,000 iterations). The results produced by all the versions were similar, except for an apparent error in the gamma distribution in Simulación. For further details see Appendix A.

The tool used for the simulations in this project was R, mainly because it was used for the analysis of the data, for which Excel was not suitable, and it was convenient to use one tool for both functions. The normal workflow was to convert the data from different sources into a database in SQLite (SQLite Consortium, 2012), which could be read directly by R for both the analysis and the simulation. R also allowed user interfaces to be written and made it simple to automate the creation of tables and graphics files for inclusion in documents. The final version of the simulation used could have been achieved in Excel, although the treatment of values less than the LoD would have been more difficult to implement. The choice reflects a preference for investing time in creating programs to automate repetitive tasks over quicker development, but more 'point-and-click', 'cut-and-paste' interaction. The use of databases rather spreadsheets to store the data was more efficient when handling large data sets and particularly when selecting subsets from them.

1.1.4 Bayesian networks and related models

Bayesian networks

Bayesian networks (also known as Bayesian belief networks or causal probability networks) are a relatively new approach to probabilistic modelling (see e.g. Jensen, 1996). In contrast to MCS, probability theory is fundamental to the way the models are constructed. A Bayesian network (BN) is represented by a network graph in which the nodes (junctions) are the variables and the arrows joining them are causal relationships (Figure 1). Thus an arrow from the *Intake* node pointing to the *Exposure* node means that *Intake* has a causal influence on *Exposure*; *Intake* is said to be a parent of *Exposure*. In most cases, each variable has only a finite number of possible values or states.

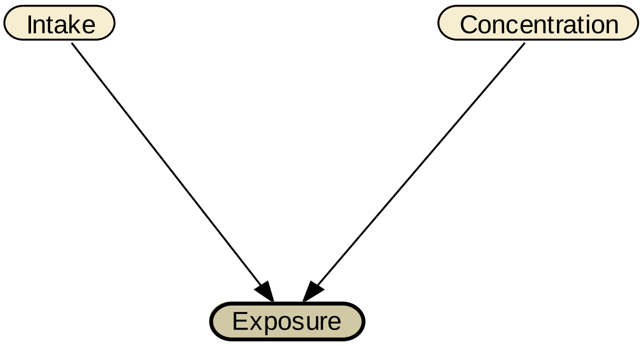


Figure 1. Simple network model for exposure, depending on intake and concentration.

Every variable has a probability distribution, that is, a probability is assigned to each possible value of the variable. A node with no parents is given an initial (or *prior*) distribution. Any other node is described by a conditional probability table, which gives the probability of the variable having each of its possible values for every combination of its parents' values. The software then calculates the joint probability distribution for the entire network and displays the distribution for each node (Figure 2). Evidence, such as observations, can be entered for any node in the network, and Bayesian inference is used to update the distributions of all the other nodes. Several software implementations are available, of which two of the most popular are Hugin™ (Hugin Expert A/S, Aalborg, Denmark) and Netica™ (Norsys Software Corporation, Vancouver, Canada).

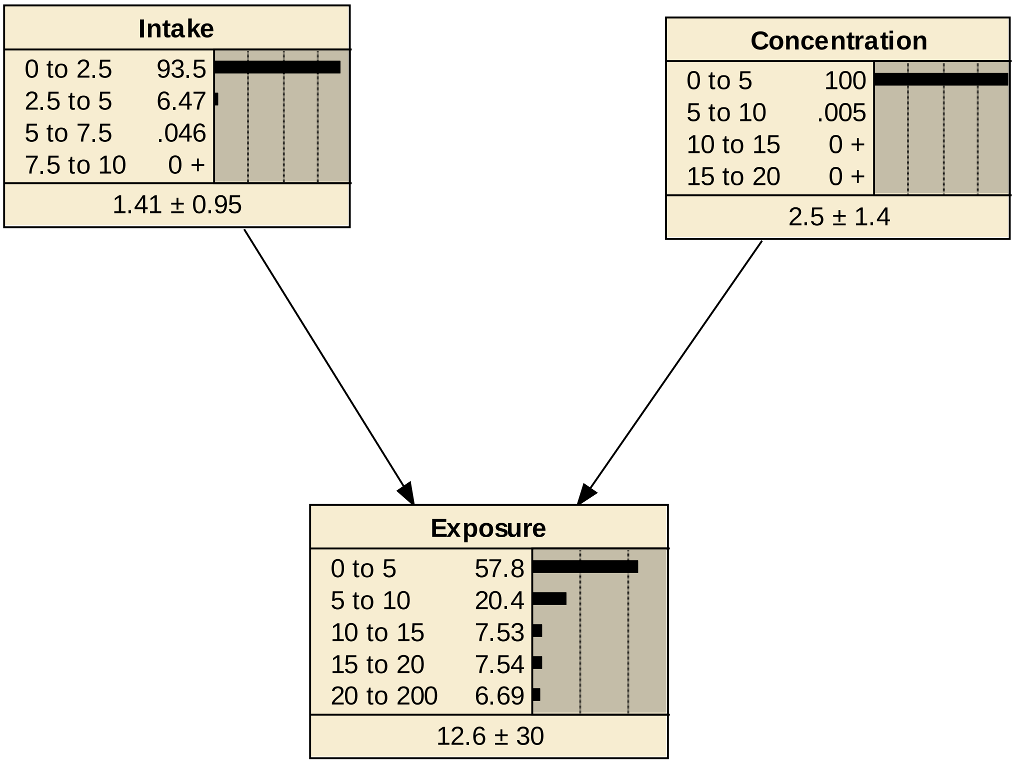


Figure 2. Bayesian network model for exposure expanded to show distributions with mean and standard error below (using a coarse discretisation).

Many of the first applications were in medical diagnosis, where the model could be constructed using the causal links from diseases to symptoms, after which inference could be used to reason from symptoms to causes. They have become widely used in many areas, including environmental modelling, especially when quantitative data or understanding of mechanistic relationships are lacking and the models must rely more on expert opinion and logical causal relationships.

The variables in Bayesian networks normally have only a small number of discrete possible states, though these can represent intervals for a numerical variable (e.g. 0–0.5, 0.5–1.0, …). This makes them less suitable for use when precise numerical results are required, because a large number of small intervals have to be used, which requires very large conditional probability tables. Netica allows continuous distributions to be specified for nodes, which it converts into sets of discrete intervals (defined by the user). It is also possible to specify the states of a child node using an equation to relate it to its parents. The conditional probability table is then generated automatically by a sampling procedure. In addition, Hugin can use continuous distributions without converting them to discrete ones, but only if the distributions are normal.

The discretization process can cause small errors. For the skewed distributions (many observations close to or equal to 0) that are typically found in risk assessment, which are also appropriate for exposure assessment, we found that these caused a slight bias, which became significant if the model contained a large number of nodes that were skewed in the same direction (Parsons *et al.*, 2005). For a simple model with few nodes, the errors would usually be negligible.

A simple exposure model was constructed as an exposure node with two parents: intake and concentration (see Figure 1 and Appendix A). The exposure was specified by an equation and the two input distributions by distribution functions. Netica then converted this to a discrete model, as outlined above. It was able to display a graph of beliefs (effectively a histogram) for each node and calculated the mean, but it was not possible to make accurate estimates of other important summary statistics, such as the upper percentiles, from the exposure histogram. (See Figure 2 for an example using a very coarse discretisation.) Netica was unable to utilise data sets such as the samples of concentration sets directly, other than by converting them into histograms, and any change to the distribution parameters required the discretisation procedure to be repeated. The strengths of the Bayesian network approach – the ability to use inference from effects to causes – were irrelevant for this application. In general, it was not well suited to the problem.

Markov Chain Monte-Carlo models

Markov Chain Monte-Carlo (MCMC) methods are a class of algorithms for generating probability distributions. One application of them has been in a freely available software package called BUGS (Bayes Using Gibbs Sampler) and now OpenBUGS (http://www.openbugs.info/w/). The program can be used to construct Bayesian network models, but solve them using simulation instead of direct inference. The advantage of this approach is that continuous distributions can be used without discretization and models (such as the product of concentration and intake) are calculated directly. Some of the features of Bayesian networks, such as inference, are available without some of the limitations or the risk of discretization errors. It also provides much of the flexibility of Monte Carlo simulation.

OpenBUGS is certainly capable of handling the exposure model: we have previously used it for a much more complex risk assessment model (Parsons *et al.*, 2005). A simple model was easily created using distributions for intake and concentration. However, the user interface was not designed for non-specialists. Furthermore, any post processing of the results would require exporting the data to other software and the graphics provided are very poor. As with Netica, the software is not especially well-suited to this problem and its advanced features, such as inference, are not required.

1.2 Approaches reviewed by IGHRC

The UK Interdepartmental Group on Health Risks from Chemicals (IGHRC) has published two relevant reviews from which we summarise the salient points.

1.2.1 Guidelines for good exposure assessment practice…

This report (IGHRC, 2004) summarised general procedures for exposure assessment and considered some alternative approaches. The steps employed in exposure assessment were:

- Problem formulation – defining the purpose, scope and level of detail.
- Data gathering – including literature review, sampling plans, exposure measurement and modelling.
- Data analysis – including statistics, identifying gaps and outliers, handling limits of detection, modelling and quality assurance.
- Exposure characterisation – summarising the estimates and evaluating their quality.

The procedures described used a generic source – pathway – receptor model. The important variables in assessing exposure were magnitude, duration and frequency. Direct exposure measurement may be preferred, when it is possible, but often it is necessary to use indirect methods, deducing exposure from concentration and intake.

The report drew on guidance for the use of descriptive and inferential statistics in data analysis (Armitage *et al.*, 2001; WHO, 2000). It considered the derivation of summary measures such as means medians percentiles and estimates of variability, or the provision of simple statistical models. Special care should be used when drawing inferences from data from limited sampling, and when dealing with skewed distributions where standard summary statistics may be inadequate to fully describe the data. Outliers should only be removed where there is strong evidence that they are erroneous. The presence of data points below the limit of detection should be managed carefully. If the focus of interest is on the upper extremes of the data set, they may have little effect, but they can cause serious difficulties in estimating the mean and standard deviation. One approach taken by the Food Standards Agency is to give two estimates: one where values below the LoD are assumed equal to 0 and one where they are assumed equal to the LoD. Quality assurance procedures should apply to the overall management and organisation of the data to ensure the reliability and reproducibility of the results.

The majority of the modelling considered by the report was deterministic. In these cases uncertainty analysis was recommended to consider uncertainty in scenarios, parameters and the model, with sensitivity analysis to examine the impact of uncertainties on the final exposure assessment. The only approach to probabilistic or stochastic modelling considered was Monte Carlo simulation using spreadsheets with plug-in software. The authors reported no consensus on the best approach of deriving distributions from data among the three outlined in section 1.1.2.

In the context of the current study, it should be noted that the exposure route has been defined clearly as ingestion of tap water, the effective source is the tap water, as the cause of contamination is outside the scope of the study, and direct exposure measurement is absent, so derivation from intake and concentration is required.

1.2.2 Current approaches to exposure modelling…

This report (IGHRC, 2010) summarises the approaches taken by three UK regulators: the Environment Agency (EA), the Health and Safety Executive (HSE) and the Food

Standards Agency (FSA). The EA is concerned with complex, multi-pathway, diffuse pollution processes resulting in human exposure through a variety of routes, but less commonly by ingestion. For this they use several different models, but all of those described were deterministic. Exposure assessment within the HSE was largely for the regulation of chemicals, including operator and consumer exposure. Again the main routes were dermal and inhalation rather than ingestion. Many of these, especially the models used for Tier 1 assessments, made worst-case assumptions, such as release of high volumes of hazardous materials within a confined space. As with the EA, the models considered in the HSE were relatively complex and deterministic.

The work of the FSA was much more relevant to the present study, since it was concerned with intake by consumers of contaminants or nutrients via food and drink. The FSA defines a five tier approach:

1. conservative calculations;
2. simple models;
3. complex deterministic models;
4. stochastic (probabilistic) models;
5. measurement.

The Tier 4 approach typically combines concentration data from random sampling of food items with consumption data provided by the National Diet and Nutrition Survey programme. An annex to the report summarises the methods used to calculate summary statistics for the population as a whole. Although it is not explicitly stated, the report implies that the Tier 4 models are Monte Carlo simulations.

1.3 Conclusions

Monte Carlo Simulation is by far the most common approach to probabilistic exposure modelling. Of the alternative methods reviewed, the analytical approach was quickly rejected as unlikely to be tractable and unsuited to data sets that are not described by parametric distributions. The two Bayesian approaches did not match the requirements of the project: they are strongest when inference is an important requirement or when data is limited.

The main feature of most of the methods reviewed by IGHRC was the complexity of the models required, rather than the emphasis on the probabilistic aspect. In fact, most of the models were deterministic, so the relevance to the present study was generally low.

The FSA Tier 4 method is directly comparable with the objectives of the present study. The data sources – the Diet and Nutrition Survey for intake and sampling for chemicals – are analogous to those available for drinking water. It also provides one approach to dealing with the issue of concentrations below the limit of detection.

2. Data sources for tap water intake

2.1 Survey sources from the UK

One of the two requirements for probabilistic exposure assessment via drinking water, in the absence of direct measurements, is suitable data on water intake. The primary sources for the UK are the drinking water or tap water consumption surveys carried out in 1978 (Hopkin & Ellis, 1980), 1995 (MEL Research, 1996) and 2008 (Accent, 2008). Although the titles of the surveys varied, we will refer to these for convenience as DWCS 1978, 1995 and 2008. The last of these is the most directly relevant for current exposure assessments, but differences between surveys, such as the inclusion of children in 1978 and 1995, but not 2008, mean that the earlier surveys may be informative. Data on water intake by adults aged 19–64 from the National Diet and Nutrition Survey (Henderson *et al.*, 2002) carried out for the Food Standards Agency (FSA) and Department of Health (DH) in 2000–01 was also made available during the project. We will refer to this as NDNS 2001. To avoid repetition, the results of the surveys are summarised together in Sections 2.6–2.9.

During the project, we also received data from a survey of intake by children less than 16 years old carried out for Defra/DWI in 2011–12 (Ipsos MORI, 2012), which we refer to as DW16 2012.

2.2 Drinking water survey 1978 (DWCS 1978)

The first large-scale survey of drinking water consumption was carried out by the Water Research Centre in 1978 (Hopkin & Ellis, 1980). Its objectives included estimating the volume of tap water consumed by individuals in Great Britain, obtaining information on the consumption of other beverages and determining what systematic variations in consumption habits existed within the population. The survey was carried out between 18th September and 20th October 1978. It covered a total of 1320 households yielding data for 2722 adults, 842 children under 15 years of age and 68 babies. It used location sampling from 100 constituencies to obtain a sample that reflected the demographic breakdown of Great Britain.

The only data from this survey available to the project were the tables and histograms in the report. The distribution of intakes had to be estimated by measuring the height of the bars in the printed report. The interval width in the histogram was 0.15 l/d.

2.3 National tap water consumption study 1995 (DWCS 1995)

The 1995 survey was commissioned by the DWI and carried out by MEL Research to update the results of the previous survey (MEL Research, 1996). It was necessary to modify the methods to take into account the increased availability of bottled water and the growth in mixed-water foods such as noodles and rice and pasta. The survey used a face-to-face questionnaire completed by the head of each household, a self-completion questionnaire completed by each member of the household and a consumption diary kept by each member of the household using standard sizes for vessels, rather than measured volumes and diary records of the proportions consumed in broad steps (e.g. 25%, 50%, 75%, 100%). The aim was to include 500 households, which were chosen by selecting 10 households from each of 50 local authority areas. The survey was completed by 476 households representing 1018 individuals

The only data from this survey available to the project were the tables and histograms in the report. The distribution of intakes had to be estimated by measuring the height of the bars in the published PDF file, which appeared to have been scanned from the original report. The interval width in the histogram was 0.25 l/d.

2.4 National tap water consumption study 2008 (DWCS 2008)

The most recent source of intake data available for the project was the national tap water consumption study (NTWCS) (Accent, 2008; Marsden, 2010), which contained self-recorded data on the intakes of about 1500 adults from 1000 households in two periods in Spring and Summer 2008. It recorded average daily intakes based on seven-day diaries using a method similar to DWCS 1995. Wave 1 in March–April contained 1527 individuals and Wave 2 in June–July contained 1446 individuals. The survey design included the use of the same households in both waves in order to compare spring and summer intakes. Of the households in Wave 1, 555 participated in Wave 2, so they were not completely independent data sets, but there was no way to identify this effect within the data.

Accent (2008) gives demographic data and some analyses stratified by the demographics. Marsden (2010) contains tables giving frequency distributions in 0.1 l/d steps for both total and unboiled consumption for both waves for the full sample, but not for sub-groups by factors such as age and sex.

2.5 National diet and nutrition survey 2001 (NDNS 2001)

NDNS 2001 (Henderson *et al.*, 2002) is one of a programme of national surveys with the aim of gathering information about the dietary habits and nutritional status of the British population. The survey covered the whole of mainland Britain and took place from July 2000 to June 2001. The sample consisted of adults aged 19–64 and excluded women who were pregnant or breastfeeding at the date of initial contact. The data were collected using seven-day self-recording diaries and included tap water. Participants were encouraged to weigh all food and drink before consumption and water added to food was estimated using a set of standard ‘recipes’. In total, 1724 people completed the diaries.

2.6 Results for the full survey samples or all adults

DWCS 1978 included adults and children. The mean tap water intake for the full sample was 0.955 l/d. From the data given in the tables, the mean for all adults was 1.084 l/d.

DWCS 1995 also included children. The mean tap water intake for the full sample was 1.138 l/d. This was reported to be significantly greater than in 1978 using a two-sided *t*-

test ($p < 0.01$). The report noted that the difference may be due to a change in the method of recording intake. According to the DWCS 2008 report, the mean for adults in DWCS 1995 was 1.275 l/d; we obtained a similar value (1.278) by extracting the data by age group from a graph in the report and excluding children.

DWCS 2008 included adults only and found mean intakes of 1.275 l/d in Wave 1 and 1.314 l/d and in Wave 2. The difference between the waves was reported not to be significant, and we found the mean for the complete set was 1.294 l/d. The report did not state whether the difference between 2008 and 1995 was statistically significant. The test of the difference between 1978 and 1995 in the 1995 report used an estimate of the standard error of the mean of 0.0129 derived from the 1978 data. We estimated the equivalent value for the 2008 data to be 0.0121. Using a two-sided *t*-test as in the 1995 report, with either estimate of the standard deviation, the difference was not significant ($p > 0.25$).

The NDNS 2001 data included adults aged 19–64 and gave individual intake records for coffee, tea and non-diluent water (which includes water in other recipes) from 7-day dietary diaries. Daily average water intakes were derived from these records. The mean intake was 1.103 l/d and the standard error of the mean was 0.0140. Using the two-sided *t*-test, the difference in the means between this survey and DWCS 1995 and 2008 was significant ($p < 0.01$). These two surveys assumed standard vessel sizes and used diary records of the broad proportions consumed, which may tend to bias the estimates of drink sizes upwards.

2.7 Analysis by sex

All of the surveys showed some difference in mean intakes by the two sexes, although these differences were not always significant (Table 1). The direction of the difference was inconsistent: in DWCS 1978 and NDNS, females drank less tap water, whereas in the other two surveys they drank more. It is interesting, although not necessarily relevant, to note that these correspond to different methods of recording intake noted above. Our analysis of variance in the NDNS data found that the difference was significant ($p < 0.01$).

Table 1. Variation of mean tap water intake by sex for DWCS and NDNS surveys.

Survey	Male, l/d	Female, l/d
DWCS 1978 all	0.980	0.933
DWCS 1978 adults	1.127	1.044
DWCS 1995 all	1.127	1.149
DWCS 2008 adults Wave 1	1.207	1.282
DWCS 2008 adults Wave 2	1.303	1.322
NDNS adults	1.140	1.073

2.8 Analysis by age

Although the surveys analysed the data by age, they all used different bands. The report on DWCS 2008 grouped the bands used in 1995 into rough equivalents of those used in 2008 to present them in a single table; here we show them separately using the original bands (Table 2-Table 4). Our analysis of the result from NDNS is included with DWCS 2008 using the same age bands (Table 4). The general pattern is consistent: intake generally increases with age, levelling off or declining from late middle-age onwards. We found that the effect in NDNS was significant ($p < 0.001$).

Table 2. Variation of mean daily tap water intake by age in DWCS 1978.

Age	Male, l/d	Female, l/d
1–4	0.477	0.464
5–11	0.550	0.533
12–17	0.805	0.725
18–30	1.006	0.991
31–54	1.201	1.091
55 and over	1.133	1.027

Table 3. Variation of mean daily tap water intake by age in DWCS 1995.

Age	Tap water, l/d
0–5	0.503
6–15	0.603
16–25	0.974
26–35	1.199
36–45	1.277
46–55	1.493
56–64	1.323
65 and over	1.382

Table 4. Variation of mean daily tap water intake by age in DWCS 2008 and NDNS.

Age	DWCS 2008 Wave 1, l/d	DWCS 2008 Wave 2, l/d	NDNS, l/d
16–24*	1.034	1.120	0.765
25–39	1.256	1.315	1.050
40–54	1.411	1.441	1.188
55+	1.322	1.447	1.189

* The lowest band in NDNS was 19–24

2.9 Other factors

The DWCS studies had also included socio-economic group and the results were summarised in DWCS 2008. The section of the table for tap water is reproduced here (Table 5). Different patterns were seen in the different surveys and the differences between groups within each were often small and not all were significant. Our analysis of variance of the NDNS data, which used a different classification, found no significant effect.

Table 5. Variation of mean daily tap water intake by socio-economic group in DWCS.

	1978, l/d	1995, l/d	2008 Wave 1, l/d	2008 Wave 2, l/d
A,B	0.881	1.104	1.306	1.371
C1	0.932	1.154	1.243	1.327
C2	0.957	1.171	1.238	1.313
D,E	1.013	1.066	1.354	1.329

The NDNS data also contained the weight and height of the subjects. DWCS 2008 reported generally increasing intake with weight (Table 6). The change of weight with age followed the same general pattern as intake, and the mean weight for males is greater than females, as was the intake in NDNS, so it appeared possible that weight was the underlying explanatory variable. When we included weight in the analysis of variance, we found no significant effect and no significant interactions with the other factors, thus age and sex appear to be the main factors influencing intake. The same was true for height.

Table 6. Variation in liquid consumption by weight in DWCS 2008.

Weight, stone*	Total liquid, l/d		Tap water, l/d		% Tap water	
	W1	W2	W1	W2	W1	W2
< 9	1.803	1.890	1.200	1.302	67	63
9 – 11	1.790	1.885	1.221	1.269	68	67
11 – 13	2.011	1.964	1.293	1.273	64	65
13 – 15	2.097	2.188	1.355	1.426	64	65
> 15	2.132	2.123	1.422	1.403	67	66

* The questionnaire recorded weigh in stones; 1 stone ≅ 6.350 kg (approximately)

2.10 Intake data for children aged 0–15 years

The survey of water intake by children here referred to as DW16 2011 (Ipsos MORI, 2012) included 1241 individuals who were aged 0–15 at the start of the survey. The intakes were recorded by the children or their parents/guardians as a 7-day on-line diary from 13th to 19th May 2011. Volumes were described using standard container sizes (e.g. cup = 200 ml, mug = 275 ml) and the proportion filled and drunk. The respondents could specify the measured volumes if they were known. Tap water and other sources were recorded separately.

The age and sex of the child was recorded for all the participants. Other information that was requested, but not always provided, included height, weight and socio-economic group. We received the data for all the participants giving the mean daily intake of tap water and total liquid, age, sex, socio-economic group, weight, height and region. Five individuals had zero tap water intakes; of the remaining 1236 the weight was recorded for 983.

Water intake was expected to vary with sex, age and weight. The effects of age and body weight were explored using linear regression and all three factors were examined by analysis of variance.

The regression of intake on age had intercept 0.395 l/d and slope 0.020 l $d^{-1}year^{-1}$, both of which were significantly different from 0 (p = 0.001), but R^2 of only 0.061 due to the very high scatter in the data. The regression of intake on body weight had 0.352 l/d and slope 0.0066 l $d^{-1}kg^{-1}$, again both significantly different from 0 (p = 0.001), but R^2 of only 0.085. We concluded that neither of these was useful for the model.

For the analysis of variance, the sample had to be divided into groups by age and weight. The groups used for age were based on the ones used in the main sources of Reference Nutrient Intake (RNI) values (Defra, 2010; Buttriss, 2000): 0–3 years, 4–6 years, 7–10 years and 11–15 years. The RNIs usually separate babies younger than 1 year from the 1–3 group, but there were insufficient babies in the sample to provide reliable results. The weight groups were constructed using the quartiles rounded to the nearest whole number to give roughly equal numbers in each group: <17 kg, 17–26 kg (*i.e.* ≥17 & <26), 26–41 kg and ≥41 kg.

The resulting means and maxima are shown in Table 7. The means were lower than for the adult population, but consistent with the youngest individuals in NDNS 2001. The effects of sex, weight and age were all significant (p = 0.001). As found in DWCS 1978 and NDNS, females drank less tap water than males. The effect of weight was more systematic than that of age, though the intake ranges were still wide and overlapping (Figure 3). There were too few babies under 1 year to model them independently of the 0–3 years group, but it may be useful to note that the mean intake for this group was 0.432 l/d and the maximum was 0.859 l/d. The mean, minimum and maximum body weights for the four weight groups were calculated for use when simulating relative exposure (Table 8).

Note that Ipsos MORI (2012) used a weighting procedure when presenting statistics for the full set, to take into account differences in the proportions in the population and the sample by age and sex. In this study, we have simply used the data from the sample, as the differences in the population statistics are generally small and the main interest here is in the results for specific sub-groups.

Table 7. Mean and maximum tap water intake for sex, age and weight groups in data from DW16 2011.

Group	Mean intake, l/d	Maximum intake, l/d
All	0.551	2.620
Females	0.511	2.085
Males	0.584	2.620
Age 0–3	0.540	1.988
Age 4–6	0.509	1.249
Age 7–10	0.585	2.085
Age 11–15	0.562	2.620
Weight < 17 kg	0.427	2.085
Weight 17–26 kg	0.466	1.839
Weight 26–41 kg	0.559	1.952
Weight ≥ 41 kg	0.705	2.620

Table 8. Group mean, minimum and maximum weights for individuals in the DW16 2011 data set.

Group	Mean, kg	Minimum, kg	Maximum, kg
Weight < 17 kg	12.95	4.99	16.78
Weight 17–26 kg	21.46	17.00	25.85
Weight 26–41 kg	32.97	26.00	40.82
Weight ≥ 41 kg	53.54	41.28	95.25

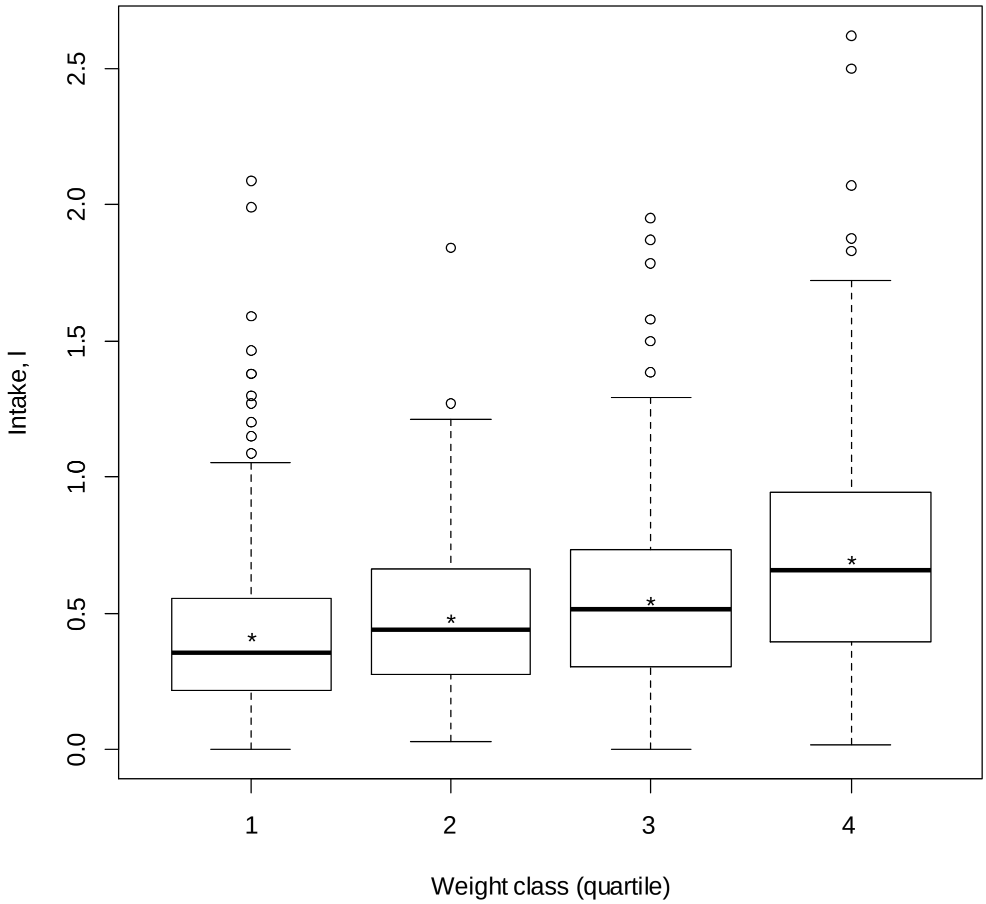


Figure 3. Box-and-whisker plot of the variation in tap water intake with weight (classified by quartiles) in DW16 2011. * = mean, bar = median, box = inter-quartile range (IQR), whiskers = median ±1.5×IQR, circles = outliers.

2.11 Comments on the UK surveys

As has already been noted, surveys that use measured volumes appear to produce lower estimates of intake than those that estimate the amount drunk based on assumed standard vessel volumes and broad proportions consumed. However, no direct comparison has been made between the two methods. There is thus no single 'best' survey to use for current intakes, as the most recent one using measurements (NDNS 2001) is over a decade old. All the surveys were used in the simulations, to allow the implications of the different estimates to be considered.

All the surveys used sampling methods that were intended to ensure that they were representative of the population, so we will assume that this was the case.

With the exception of NDNS 2001, the data available were weekly means of daily intake. The use of weekly means underestimates the variability of daily intake. Some element of smoothing is desirable, for example to reduce artefacts due to drinks taken just before or just after midnight. If the aim of the study was to look at acute toxicity, the daily extremes would be more important, but this is not the case with the substances being considered. Although the raw intakes were available for NDNS 2001, the survey used a one week period and was designed to estimate the intake over a week, not the daily variability, so the data were averaged for this study to be comparable with the other survey results.

There survey results provide no data on how the mean daily intake by individuals varies over time, for example whether some are persistently low or high consumers, which seems likely. For this study, the data points are treated as independent daily intake events.

2.12 Other surveys of drinking water intake

A comprehensive search strategy was developed to identify literature reporting drinking water survey findings. A detailed set of search terms were developed as the basis for identifying articles and reviews from authoritative sources such as the Drinking Water Inspectorate/Department for Environment Food and Rural Affairs (DWI/Defra), and from peer-reviewed literature, through searches of SCOPUS (includes Medline & Embase) and CSA Illumina (Aqualine, Biological Sciences, Environment Abstracts, Environment Science and Pollution Management, Medline, Risk Abstracts, Toxline, Water Resources Abstracts).

The output of these searches was subjected to detailed consideration for relevance to the project by a risk assessor within Cranfield University's Institute of Environment and Health.

2.12.1 Sweden (Westrell *et al.*, 2006)

Background

Estimates on drinking water consumption are necessary in risk assessments on microbial hazards in drinking water. Although point estimates can be applied in risk assessments it is preferable to use statistical distributions of parameters which account for variability and uncertainty in the data. There appear to be large differences in consumption habits of tap water between countries and for demographic variables such as age, sex and geographic area.

Aims

The aims of this study were:

- to make a quantitative estimation of drinking water consumption in Sweden; and
- to evaluate potential differences in demographic variables that could impact on water intake

Methods

Data used in the study originated from:

- National Environmental Health Survey (NMHE99) undertaken in Sweden in 1999, relating to the physical health of the Swedish population. Cohort size related to daily water intake was 10,957. Questionnaires were sent out to Swedish people living in 21 countries asking for estimated water intake, separated into cold tap water consumed as drinking water, tap water prepared into coffee, tea etc. and daily bottled water consumption. Responses could be either 1 l or more, less than 1 l, or none.
- Data collected from an investigation of a waterborne contaminant outbreak in Transtand in February – March 2002. Questionnaires were sent out and received from 157 permanent residents in the area of the outbreak, asking whether tap water had been consumed at home or elsewhere and number of glasses (one glass being 200 ml) of unboiled tap water normally consumed in one day.
- A small heated tap water and bottled water consumption study in which questionnaires on water consumption and consumption patterns were obtained from 75 Swedish residents. The amount of heated tap water consumed within and outside the home as pure water, or as a base for tea, coffee, chocolate, soup etc. and the amount of bottled water consumed per week (responses of none, 0.5 l, 0.5–1.0 l, 1.01–2.0 l, 2.01 l or more) was assessed.

Results

A lognormal distribution was fitted to the daily direct/cold water intake from the Transtand survey with $\mu = -0.299$ and $\sigma = 0.570$ (the parameters of the normal distribution after the log transform). The average daily consumption of tap water as plain drinking water was 0.86 ± 0.48 l/d and as heated tap water, e.g. in coffee and tea, was 0.94 ± 0.69 l/d, so the mean total tap water intake was 1.8 l/d. Women consumed more cold tap water than did men ($p < 0.001$), while men appeared to have a higher consumption of heated tap water. Cold tap water intake was significantly different between age groups ($p < 0.001$) with highest intake in the oldest age group (≥ 70 years). There was also a significant decrease in cold tap water intake with increasing yearly income ($p < 0.001$, correlation coefficient −0.091). The consumption of bottled water was very low (mean 0.06 l/d) when compared to other countries.

The intake of heated tap water in beverages was less than the amount of tap water consumed directly and no difference in the intake of heated tap water between sexes or ages were seen ($p = 0.773$ and 0.180 respectively). A negative correlation was noted between increasing yearly income and amounts of tap water consumed in beverages ($p < 0.001$, correlation coefficient −0.057).

Discussion

The authors highlight several limitations within the study:

- Use of questionnaires on consumption on a single day (may overestimate intake).

- Questionnaire based on consumption at home (may be biased by sex, age).
- Employment status may be a stronger determinant of tap water consumption patterns than sex.
- Although children (0-9 and 10-19 years) had the lowest consumption, ingestion in terms of body weight will be highest (approximately 3-4 times higher for babies less than 1 year than the general population).

The sample size used in the quantitative analysis was small (157) and the use of a fixed unit of intake (200 ml glass) could have led to errors in estimation.

2.12.2 Germany (Sichert-Hellert *et al.*, 2001)

Background

Few studies have been carried out to look specifically at water intake and beverage consumption in healthy children and adolescents. However, children have a higher body water turnover and subsequently, a higher water requirement than adults.

Aims

To use data collated during the continuing DONALD Study (Dortmund Nutritional and Anthropometric Longitudinally Designed Study) to evaluate water intake and beverage consumption in 2–13y-old subjects.

Methods

Water intake was evaluated and time trends in water intake and beverage consumption were assessed on the basis of 3 d weighed dietary records (n= 3736) of 2–13 year old males (n= 354) and females (n= 379) enrolled in the DONALD Study.

Results

Total water intake increased with age from 1114 g/d in the 2-3-y-olds to 1363 g/d in the 4–8 year olds and further to 1801 g/d (1676 g/d) in the 9–13 year old boys (girls); 33–38% came from food, 49–55% from beverages and 12–13% from fat oxidation. Total water intake relative to body weight decreased with age from 77.5 g/kg (boys and girls) to 48.9 g/kg in boys and 42.6 in girls. Milk (9–17%) and mineral water (12–15%) were the most important sources of total water intake. In the 15 year period a significant increase in total water intake (+1.7 to +3.2 g MJ^{-1} y^{-1}) in all three age groups irrespective of sex was found. The increase in total water intake was mainly due to an increase in beverage consumption (+0.32 to +0.47%/y).

Discussion

The comparison of this data with other surveys points to a low total water intake, especially a low tap water intake, in German children and adolescents and underlines cultural influences on food and drinking habits.

The authors highlighted some points that make comparisons between studies difficult:

- A lack of clear definitions of beverage items.
- Errors due to semi-quantitative measures (cup, can).
- Errors due to measuring total weight (including nutrients) instead of water intake only.

2.12.3 Germany (Hilbig *et al.*, 2002)

Background

During the first 4-6 months of life, non-breast fed infants are dependent on tap water for the preparation of milk formula. Contaminated tap water (e.g. with metals, or environmental pollutants) can become a health risk in sensitive populations such as infants and young children. However, there is a lack of data on measured water intake in normally nourished healthy infants and children.

Aims

To use individual food and fluid intake measurements from healthy infants, children and adolescents in the DONALD (Dortmund Nutritional and Anthropometric Longitudinally Designed) Study, to report on the distribution of individual intakes of tap water in infants and young children aged 3–36 months.

Methods

In the DONALD Study, food consumption is assessed by 3-day weighed diet records, with all foods and fluids (including tap and mineral water) being weighed to the nearest 1g. Measurements are taken at 3, 6, 9, 12, 18, 24 and 36 months of age. This study utilised 1962 diet records from 504 subjects between 1990 and 1998 to calculate scenarios for potential tap water contamination.

Results

Tap water intake relative to body weight was significantly higher in formula-fed (FF) infants than in breast-fed (BF) infants under 1year ($p<0.0001$). The estimated median intake of lead and nitrate relative to body weight from tap water was higher in FF infants than in BF infants or mixed fed (MF) young children. The scenarios based on intakes at the median, 95th percentile or maxima show that higher risks for exceeding the presently existing maxima could be expected in FF infants.

Discussion

The authors conclude that their study used readily available data to show potential risks to exceed tolerable values for contaminants in drinking water e.g. lead and nitrate.

2.12.4 France (Gofti-Laroche *et al.*, 2001)

NB: The article is written in French; the information below was taken from the abstract.

Background

Assessment of risks associated with waterborne pollutants requires a good characterization of the exposure of individuals and populations. This characterization implies knowledge of pollutant levels in water and their temporal variability, along with an estimation of drinking water consumption.

Methods

This work, included within the E.M.I.R.A study which was set up to assess waterborne infectious risks, describes in details daily drinking water consumption of 544 French volunteers. Data were collected by self-questionnaires.

Results

Results differ according to the season. Tap water usage for food follows a normal distribution (arithmetic mean in winter=1.55 l/d, 95% CI [0.20-2.90]; arithmetic mean in spring=1.78 l/d, [0.13-3.43]). Total drinking water intake follows a lognormal distribution (geometric mean in winter=1.60 l/d, standard deviation=1.73 l/d; geometric mean in spring=1.92 l/d, standard deviation=1.70 l/d). Tap water intake amounts to more than 80% of total drinking water consumption, and pure tap water (i.e. not added, modified nor boiled) amounts to 42% of total drinking water.

From the information given, the lognormal distribution for total drinking water intakes have $\mu = 0.47$, $\sigma=0.55$ in winter, with a mean of 1.8 l/d, and $\mu = 0.65$, $\sigma=0.53$ in spring, with a mean of 2.2 l/d.

2.12.5 USA (USDA, 2011)

Background/Aims

The goals of this study were to describe plain drinking water intake patterns of the U.S. population and determine whether total, tap, and bottled water intakes differ by gender, race/ethnicity, income, and activity level.

Methods

Twenty-four-hour dietary recall data from 16,566 individuals age 2 years and over participating in What We Eat In America (WWEIA), the dietary intake component of the National Health and Nutrition Examination Survey (NHANES), in 2005-2008 were analysed. Appropriate sample weights were applied to produce nationally representative estimates. Differences in the intakes of total plain, tap, and bottled water intake by gender, race/ethnicity, activity level, and income were identified using *t*-tests. Regression procedures were used to adjust estimated means for confounding variables when testing for differences in daily plain water intake by race/ethnicity, activity level, and income.

Results

On any given day, 76% of individuals aged 2 years and over reported consumption of plain drinking water. The mean intake per person (including both reporters and nonreporters) was 3.9 cups (0.924 l). Total plain drinking water intakes do not differ by gender within age group, but tap water intakes are higher for males 12–19 years than for females the same age and for females 60+ years than for males the same age ($p < 0.001$). The majority of plain drinking water is consumed at home. There are some differences in intakes by race/ethnicity, income, and activity level. In some age groups, including adults 20+ years, tap water intake is higher for non-Hispanic whites than for non-Hispanic blacks and Hispanics. Among adults over 20+ years, there is a positive association between bottled water intake and income, though intakes of total plain and tap water do not differ by income. Adults who are physically active drink more plain water than sedentary adults do.

2.12.6 Canada (Levallois *et al.*, 1998)

Background

Canadian average consumption of water used to derive drinking water guidelines is assumed to be 1.5 l/d, which is based on a drinking water survey of 970 Canadians in 1977/1978. This figure may not be as relevant to today's population who consume more bottled water and carbonated drinks than would have been the case during the period of the survey.

Aims

To carry out a pilot study on drinking water consumption in a sample of residents in Quebec City, with specific focus on age, sex and residential area.

Methods

A pilot study on water consumption was carried out in the Québec City region in April and May 1996 with 125 people using a 24-h recall plus a 2-day diary. Consumption of drinking water via liquid and food was assessed as well as the type of water consumed (tap, bottle or filtered water) and place of consumption (home or away from home). For each food and drink selected, the amount of water used in its preparation was calculated; some foods were considered to be 100% added water (e.g. coffee, tea and noodles) and soup 50% water.

Results

Most of the people (56%) were drinking some bottled water or filtered tap water and 25% of water intake was away from home. Food consumption was found to be a non-significant source of drinking-water intake. The average water consumption was similar in exclusively tap water consumers and bottled or filtered water consumers (1.5 vs. 1.7 l/d, $p = 0.29$. No significant differences in amounts consumed were found according to age, but older people drank hot beverages and soup more often. Estimated water consumption differed significantly between days ($p<0.05$). Highest estimated levels were found during recall, however levels also differed significantly between diary estimates on day 1 and day 2 ($p<0.05$).

Discussion
The authors identified the following weaknesses within the study:

- The present pilot-study was weakened by a low participation rate (14%). Incentive might be necessary to improve participation rate and data collection methods must also be simplified.
- A 24-h recall plus a 1-day diary seem sufficient and data on consumption could be limited to liquids, soups and cereals.

2.12.7 Other: Netherlands, Australia, Germany, UK (Mons *et al.*, 2007)

Background
This paper presents the findings of a report (Mons *et al.*, 2005) written as part of the Microrisk (Microbiological risk assessment: a scientific basis for managing drinking water safety from source to tap) project. It presents a review of tap water consumption studies. Study design was evaluated including factors that may influence consumption.

Methods
Raw consumption data were obtained from 4 countries, collected with different study designs. Statistical models were fitted to the data to determine variability in drinking water consumption and the impact of study design on outcome and statistical distribution.

Results
From the studies evaluated, the mean consumption of cold tap water for the 'average consumer' was 0.10–1.55 l/d (most studies < 1 l/d), of total tap water 0.955–2.58 l/d (only one study > 1.95 l/d) and of total water 1.14–2.19 l/d. No conclusions could be drawn regarding the effects of season, age and gender on tap water consumption. Physical activity, yearly income and perceived health status were reported to influence water consumption.

Discussion
The authors made several recommendations:

- Estimation of drinking water consumption is higher in questionnaires than in diaries; therefore diaries are the recommended choice for collecting such data. In addition, the longer the period of collection, the more representative the data is (3-4 days optimal).
- If diaries can't be used then 24 hour recall is the method of choice with repetition on one non-consecutive day to estimate within person variation.
- Large number of respondents should be recruited with collection of data spread evenly over one or more years to avoid generalisation in time; 2000 adults as minimum sample size is recommended.

- Water consumption data is best collected as continuous data (grams or litres per day) rather than in discrete units (e.g. cups) for statistical analysis.
- The Poisson distribution performed better than the lognormal distribution with the datasets, but the Poisson distribution is only suitable for discrete data not continuous measurements.
- Non-consumers can form a high proportion of those recruited; it is advised that statistical probability distributions are fitted to the total dataset, including non-consumers.

2.12.8 USA (secondary sources: Burmaster, 1998)

Background

The author's aim was to fit distributions of intake suitable for use in health risk assessments to existing data from a survey conducted in 1978 (Ershow *et al.*, 1993) for pregnant and lactating women.

Methods

The author used probability plots and maximum likelihood estimation (MLE) to fit lognormal distributions for daily intake of total water and tap water by 3 groups of women (6201 controls, 188 pregnant, and 77 lactating; all 15–49 years of age) in the United States. The data were only available as summary statistics and frequency distributions (histograms): the individual records were not used. He also developed bivariate lognormal distributions for the joint distribution of water ingestion and body weight for these 3 groups.

Results

The mean tap water intakes and lognormal distribution parameters were:

- Control: 1.157 l/d, $\mu = -0.0018$, $\sigma = 0.593$
- Pregnant: 1.189 l/d, $\mu = -0.0038$, $\sigma = 0.640$
- Lactating: 1.31 l/d, $\mu = 0.165$, $\sigma = 0.492$

Conclusions

The author recommended the distributions for water intake as fitted by MLE for use in human health risk assessments.

Comment

Distributions other than the lognormal were not considered.

2.12.9 USA (secondary sources: Roseberry & Burmaster, 1992)

Abstract

"We fit lognormal distributions to data collected in a national survey for both total water intake and tap water intake by children and adults for these age groups in years:

0 < age < 1; 1 ≤ age < 11; 11 ≤ age < 20; 20 ≤ age < 65; 65 ≤ age; and all people in the survey taken as a single group. These distributions are suitable for use in public health risk assessments."

Comment

We have not been able to obtain a copy of this paper, but assume that it used data from the same survey as the later paper (Burmaster, 1998).

2.12.10 Discussion

The mean tap water consumption reported by several of these surveys was 1.5–1.8 l/d, which is higher than the amounts found in the British surveys. The use of recall rather than diaries has been suggested to cause for the overestimation (Mons *et al.*, 2007). The use of fixed volume units rather than measured (or estimated) volumes could also be a cause.

Several of the surveys reported a trend of consumption increasing with age, as was found in Britain. However, there was no consistent effect of sex on intake.

The distribution most commonly fitted to the data was the lognormal. However, this often appeared to be an *a priori* decision, not the result of evaluating different distributions.

Table 9. Summary of drinking water consumption papers.

r (date)	Country	Participants	Method	Factors	Distribution	Mean intak
ll et al	Sweden	157 adults and children (10,957 adults and children in earlier national survey)	Recall; fixed 200 ml units	Sex, age, income	Lognormal (μ , σ) Cold water (–0.299, 0.570)	Cold: 0.86 Heated: 0.94 Total: 1.8
-Hellert 001)	Germany	733 children (2-13)	DONALD: longitudinal; semi-quantitative records	Sex, age, results standardised by weight	None	Total: 1.1–1
(2002)	Germany	504 infants (<36 months)	DONALD; ditto	Age, post-natal feeding, results standardised by weight	None	n/a
aroche 001)	France	544 adults	Questionnaire	Season	Lognormal (μ , σ) winter (0.47, 0.55) spring (0.65, 0.53)	Total Winter: 1.8 Spring 2.2
ES 2011	USA	16,566 adults and children	Recall	Sex, age, ethnicity, income, activity level	None	Cold: 0.924
is et al	Canada	125 adults	Recall + 2 day diary	Age	None	1.5 (tap wat consumers) 1.7 (bottled/filt water consu
t al	NL, Australia, Germany, UK	Review of other studies	Various; recommends diary	Various	Poission if using discrete measures	0.955–2.58. one result >
ster	USA	Women	Recall	Pregnant, lactacting	Lognormal (μ, σ) Control: (-0.0018, 0.593) Pregnant: (-0.0038, 0.640) Lactating: (0.165, 0.492)	Control: 1.1 Pregnant: 1. Lactating: 1.
rry & ster	USA	Adults and children	Recall	Age	Lognormal	not availabl

3. Modelling tap water intake

As most of the intake data sets were only available as frequency distributions, the preferred method of using them in the exposure simulation was by fitting parametric distributions to them. Of the data initially available, DWCS 2008 had the best resolution, being presented as a table of frequencies of width 0.1 l/d, so the exploratory analyses were carried out on this set and the methods were subsequently applied to the other DWCS sets and finally to NDNS 2001, which required a slightly different approach because it contained individual records.

3.1 DWCS 2008

Some of the analysis of DWCS 1978 had assumed a normal distribution for the intake (Hopkin & Ellis, 1980), but it was immediately clear from the histograms that this assumption did not hold for the 2008 data, as both total and unboiled intakes were positively skewed (Figure 4). A review of cold water consumption for microbial risk assessment (Mons *et al.*, 2007) considered data collected from several countries and tested how well several distributions fitted the detailed data sets from the Netherlands, UK, Germany and Australia. An important distinction was made between discrete and continuous data sets. Discrete data sets measure the intake in fixed units, such as glasses, whereas continuous ones (e.g. Accent, 2008) use, in principle, precise measurements of volume, though often these are estimates based on broad proportions consumed (eg 25%, 50%, 75% and 100%) and standard assumptions about vessel volumes. Some of the data sets showed consumption patterns similar to the unboiled intake in Figure 4, which were usually best fitted by an exponential distribution. Mons *et al*. made no recommendation for continuous data sets, but recommended the Poisson distribution for discrete data.

None of the British surveys used fixed volumes, so all should be treated as continuous data. Standard continuous distributions suitable for skewed data include exponential, lognormal, gamma and Weibull (see Glossary). All of these except the Weibull distribution had been tested by Mons *et al*. for the discrete data sets and all gave good results in some cases. The exponential distribution has its maximum at 0 and decreases monotonically, so it is only suitable for data similar to the unboiled intake in Figure 4. As the contaminants of interest in this project are not generally removed by boiling, it was the total intake that was of interest, so the exponential distribution was not appropriate.

In addition to the distributions above, one additional option was considered. If a random variable has a Poisson distribution, its square root is approximately normally distributed with variance 0.25 (McCullagh & Nelder, 1989). As the Poisson distribution had been found to fit discrete data, normal distributions were fitted to the square root of the intake, which will be referred to here as the *normal (square root)* or *normal (sqrt)* distribution.

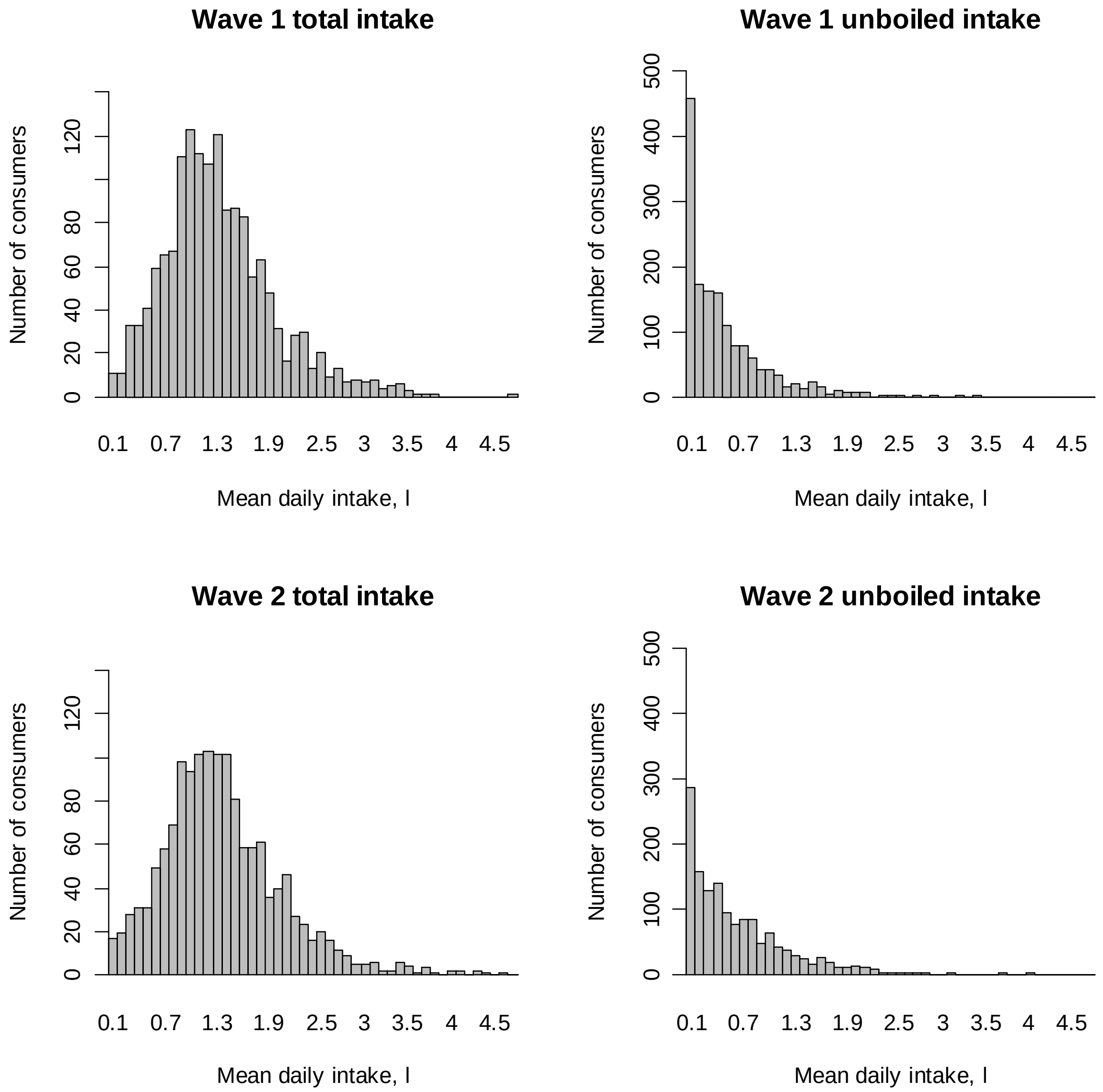


Figure 4. Distributions of total and unboiled water consumption from both waves of DWCS 2008.

The fitting functions required the data as sets of points, so it was necessary to generate them from the frequency tables. In the initial analyses, an appropriate number of points from each class in the frequency table were generated by sampling from a uniform distribution over the range of the class. This was intended to introduce some uncertainty into the values, rather than using the midpoint of each class. The data for each wave in the survey were considered individually and combined, as no significant difference had been found between the means of the two waves (Accent, 2008). The participants who were included in both waves were therefore represented twice in the combined set, but there was no information available about the consistency in the intake of the same individual at different dates.

The distributions were fitted using maximum likelihood estimation (see Glossary) via the *fitdistr* function of the *MASS* package for R. The number of degrees of freedom was constant because each distribution had two parameters: the mean and standard deviation (possibly transformed) for those related to the normal distribution, shape and rate for the gamma distribution and shape and scale for the Weibull distribution. For each, three graphical tests were used: a direct comparison of the density function with the histogram, a quantile-quantile (Q-Q) plot and a direct comparison of the cumulative distribution function (CDF – see Glossary) with the empirical cumulative distribution from the data. The Q-Q plot shows the value of each quantile of the data against that of the fitted distribution; for a perfect fit it would be a straight line of slope 1.

The Anderson-Darling (Anderson & Darling, 1952) test for goodness of fit was also applied. This is a non-parametric test based on the empirical distribution function. It computes a test statistic related to the distance between the fitted and empirical cumulative distribution functions. Compared to the more familiar Kolmogorov-Smirnov test it gives more weight to the tails of the distribution and is also more appropriate when distributions are being fitted to the data. The upper tail, quadratic class version (roughly equivalent to using mean-square errors) for left-truncated data (Chernobai *et al.*, 2005) was used, provided by the *ad2up.test* function of the *truncgof* package for R. The Anderson-Darling tests were repeated five times, because the simulation used in the Anderson-Darling method resulted in variability in the p value. (Note: since the project was completed, the truncgof package has been withdrawn from R due to possible incompatibility with recent versions.)

In practice, the test was very sensitive and not always informative. The p values could not always be calculated and were low even when data were generated artificially from known distributions. A cruder comparison was therefore made by calculating the residual sum of squares (RSS) between the fitted and empirical cumulative distribution functions.

The results are shown in Table 10 and Figure 5–Figure 8. Note that the best fit is where log likelihood is highest (for example –1393 > –1565). The log likelihood is the sum of log likelihood values for all the points in the data set, so it cannot be directly compared between waves or between a single wave and the combined data because they contain different numbers of points. Note also that the log likelihood for the parameters of the normal (square root) distribution has been corrected for the effect of the square root transformation (see Appendix B). It was immediately clear from the graphs and the RSS that the lognormal distribution fitted this data set very poorly. It had a consistently lower likelihood than the other distributions. The same was found in initial tests with the other DWCS data, so we rejected it from further consideration, despite its popularity with some of the sources reviewed above. The other three distributions were all plausible and the normal (square root) was always the best fit, though the differences in the statistics were relatively small. The RSS reflected the visual assessment better than the more formal Anderson-Darling statistic.

Table 10. Log likelihood and Anderson-Darling test *p* value for distributions fitted to total water intake from DWCS 2008.

Data set	Distribution	Corrected log likelihood	Anderson-Darling *p* value*	RSS of CDF
Wave 1	Lognormal	−1565	0–0.02	0.0646
	Normal (sqrt)	−1393	0.07–0.13	0.0078
	Weibull	−1406	0–0.02	0.0124
	Gamma	−1422	0	0.0138
Wave 2	Lognormal	−1658	0.01–0.03	0.1169
	Normal (sqrt)	−1413	0.05–0.12	0.0077
	Weibull	−1431	0.01–0.03	0.0136
	Gamma	−1464	0	0.0251
Combined	Lognormal	−3155	0–0.02	0.0738
	Normal (sqrt)	−2811	0.05–0.09	0.0078
	Weibull	−2844	0.01–0.02	0.0137
	Gamma	−2861	0	0.0140

* Range when test done five times; 0 indicates that a *p* value of $< 2.2\times10^{-16}$ was returned

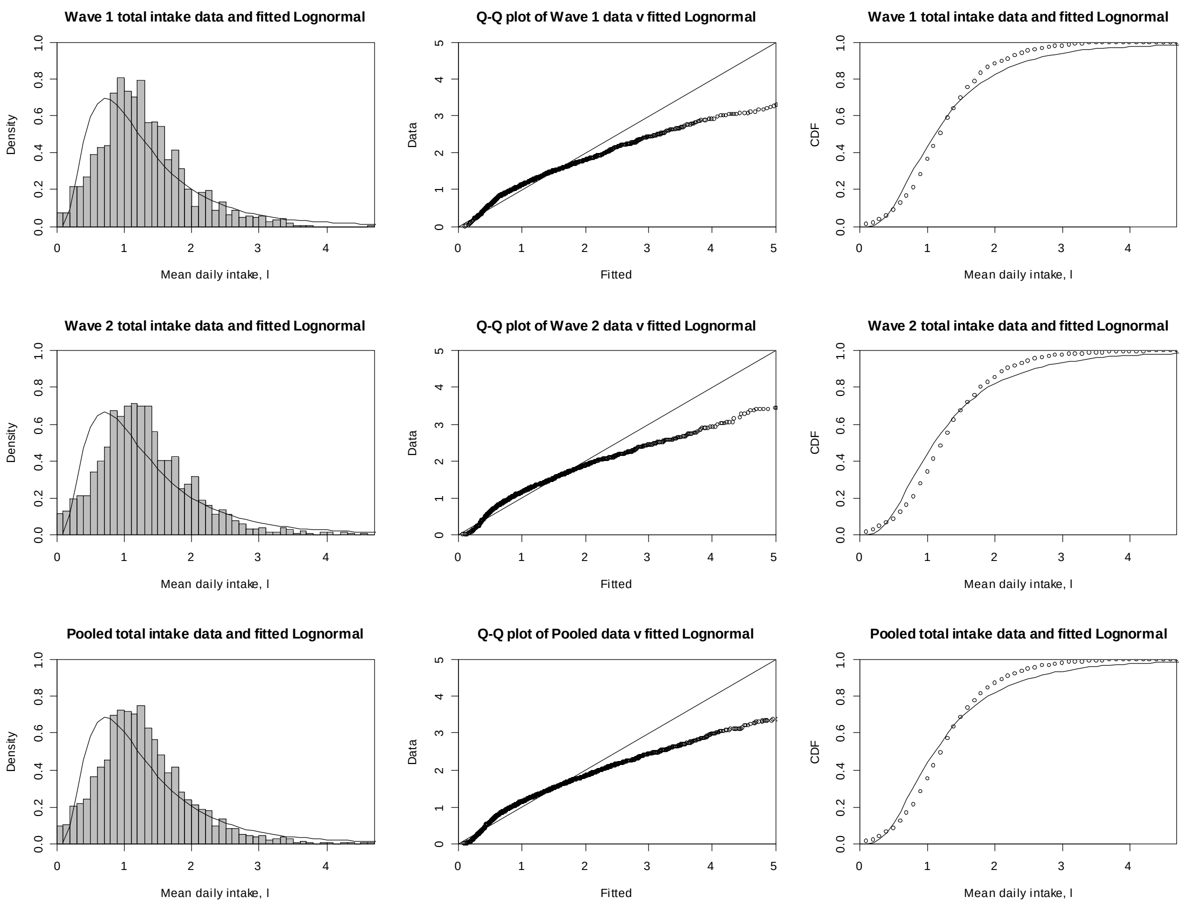


Figure 5. Lognormal distribution fitted to total intake data from DWCS 2008 for Wave 1, Wave 2 and the combined data. Graphs show the histogram and density function; quantile-quantile plot; empirical and fitted cumulative distribution functions.

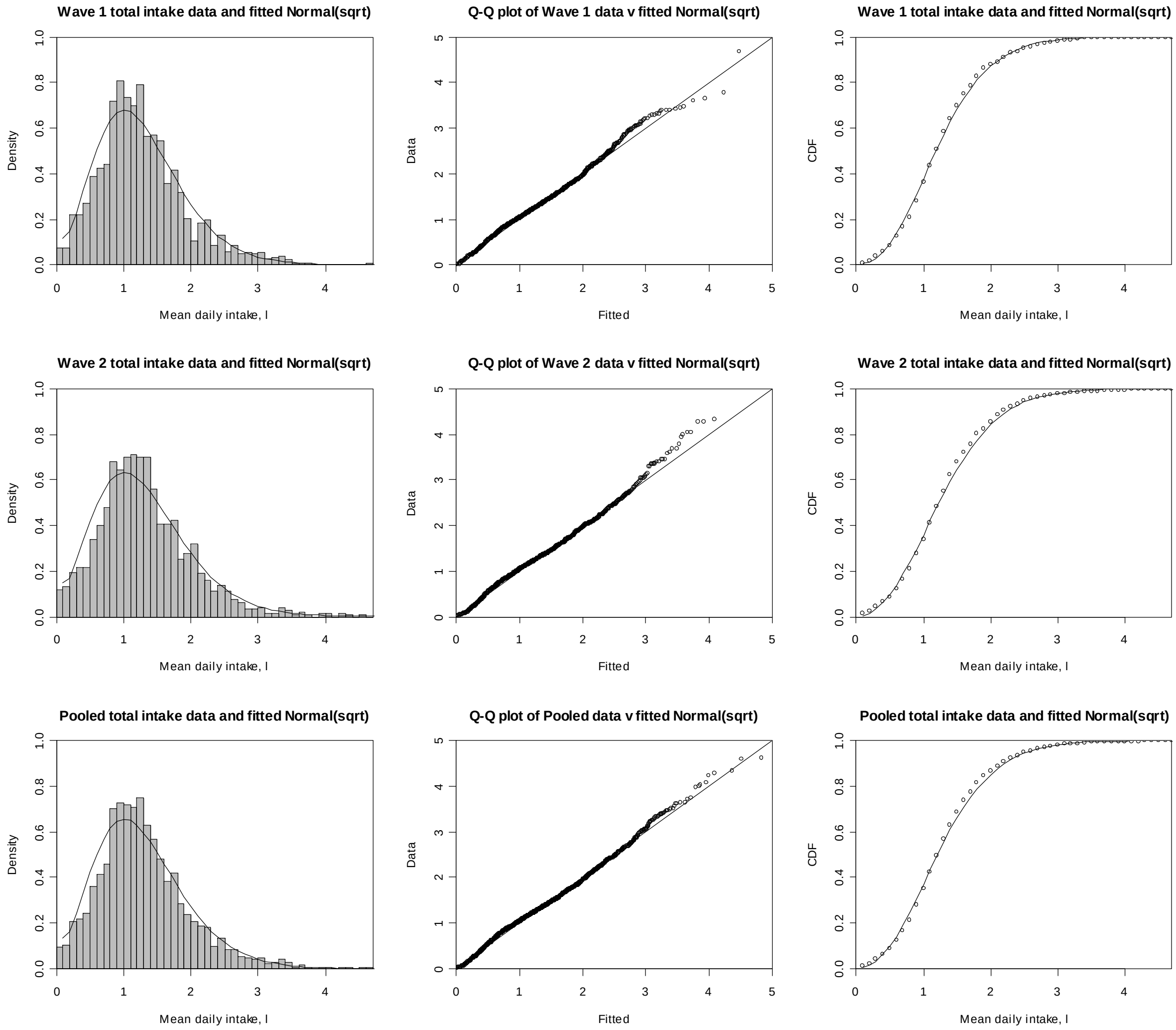


Figure 6. Normal distribution fitted to square root of total intake data from DWCS 2008 for Wave 1, Wave 2 and the combined data. Graphs show the histogram and density function; quantile-quantile plot; empirical and fitted cumulative distribution functions.

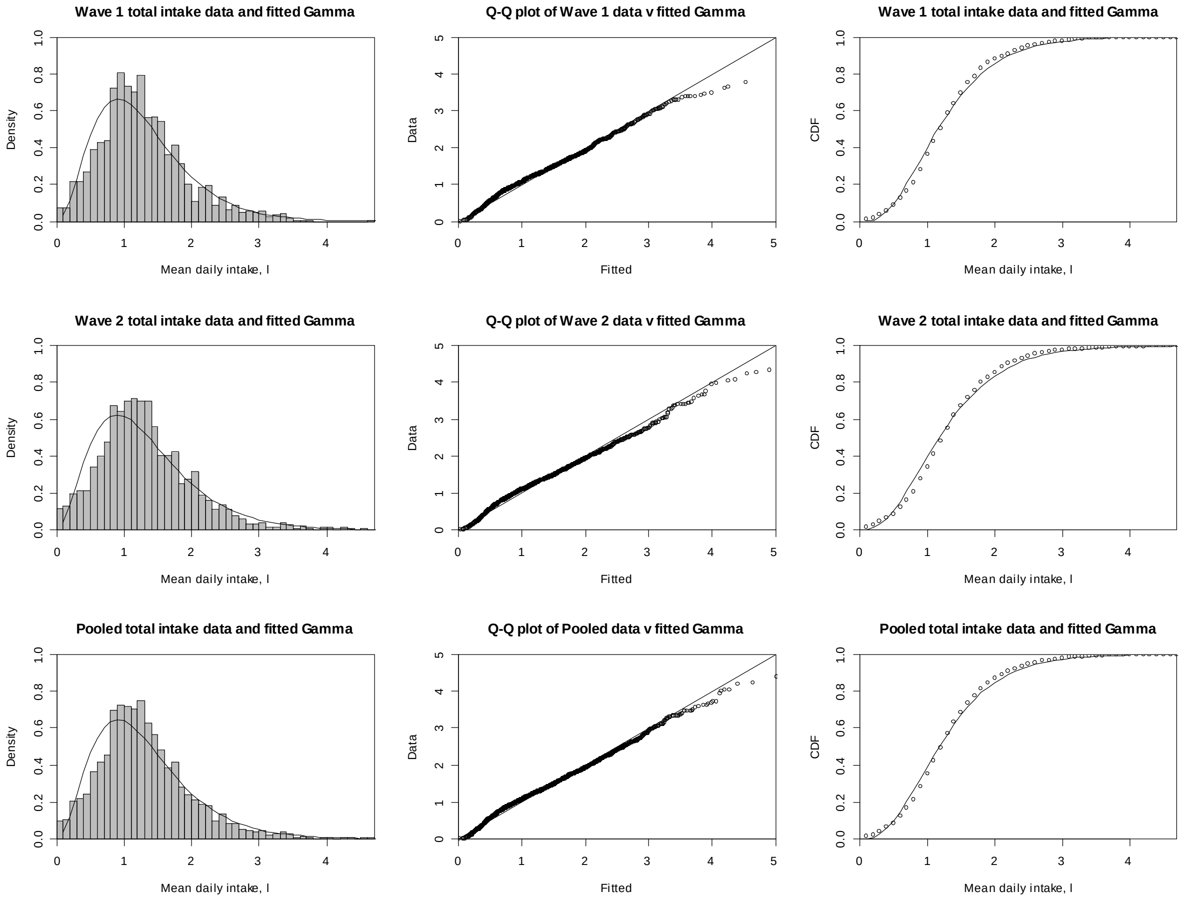


Figure 7. Gamma distribution fitted to total intake data from DWCS 2008 for Wave 1, Wave 2 and the combined data. Graphs show the histogram and density function; quantile-quantile plot; empirical and fitted cumulative distribution functions.

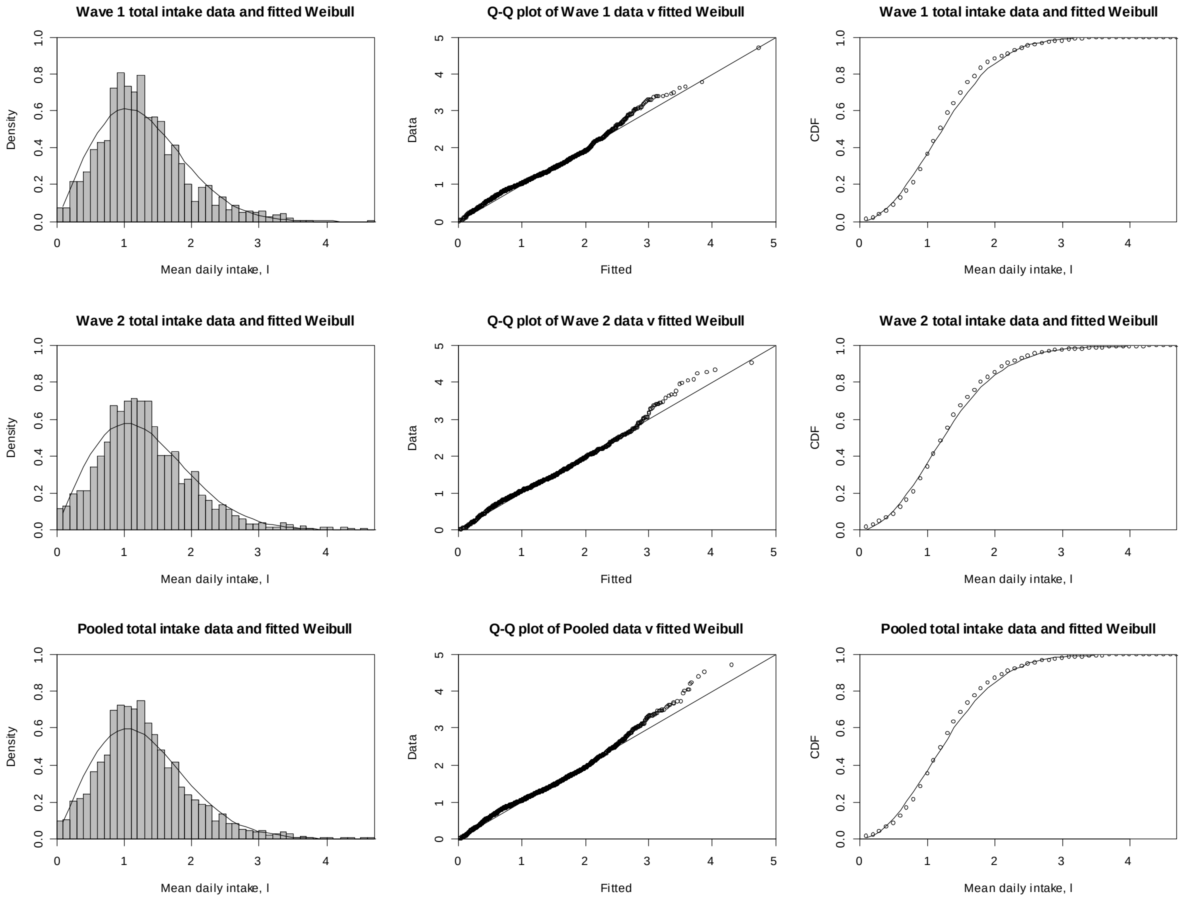


Figure 8. Weibull distribution fitted to total intake data from DWCS 2008 for Wave 1, Wave 2 and the combined data. Graphs show the histogram and density function; quantile-quantile plot; empirical and fitted cumulative distribution functions.

After the initial exploration of the data, an improved method of fitting to the frequency classes was used. The *fitdistcens* function from the *fitdistrplus* package in R is designed to apply maximum likelihood estimation to 'censored' data (see Glossary), including the case of interval censored data in which only lower and upper bounds on each value are known. To use this, each frequency class was converted into the appropriate number of points, each given by the bounds on the class value, thus eliminating any prior assumption about the location of the points within the class.

Other than the likelihood, the only goodness of fit statistics available using this function were the Akaike and Bayes information criteria (AIC and BIC), which are both derived from the likelihood and give no additional information when the number of parameters is the same for all the models being fitted. As the representation of the upper tail of the distribution is likely to be important for exposure assessment, the 99^{th} percentile of the distribution was compared with the estimate from the data.

The results of fitting the three distributions that gave acceptable results with the previous method are shown in Table 11 and an example for the total of the two waves is shown in Figure 9. Note that the empirical cumulative distribution function is now plotted with error bars derived from the uncertainty in the data values. This example for the total of the two waves illustrates the difficulty of selecting a single statistic to judge the goodness of fit. The log likelihood for the normal (sqrt) distribution was higher than for the other distributions in each case. (In this case, no correction to the likelihood value is required: see Appendix B.) In terms of predicting the 99th percentile value, the gamma distribution was the best, though the histogram/density plot shows that its peak was skewed slightly too far to the left. The Q-Q plot showed that the Weibull distribution performed poorly in the tail, which was reflected in consistent underestimation of the 99th percentile.

Table 11. Log likelihood and 99th percentile for distributions fitted to total water intake from DWCS 2008 using *fitdistcens*.

Data set	Distribution	Log likelihood	99th percentile of data	99th percentile of distribution
Wave 1	Normal (sqrt)	-4905	3.25	3.09
	Weibull	-4920	3.25	2.99
	Gamma	-4928	3.25	3.32
Wave 2	Normal (sqrt)	-4739	3.45	3.28
	Weibull	-4752	3.45	3.17
	Gamma	-4774	3.45	3.57
Combined	Normal (sqrt)	-9647	3.35	3.18
	Weibull	-9675	3.35	3.08
	Gamma	-9705	3.35	3.44

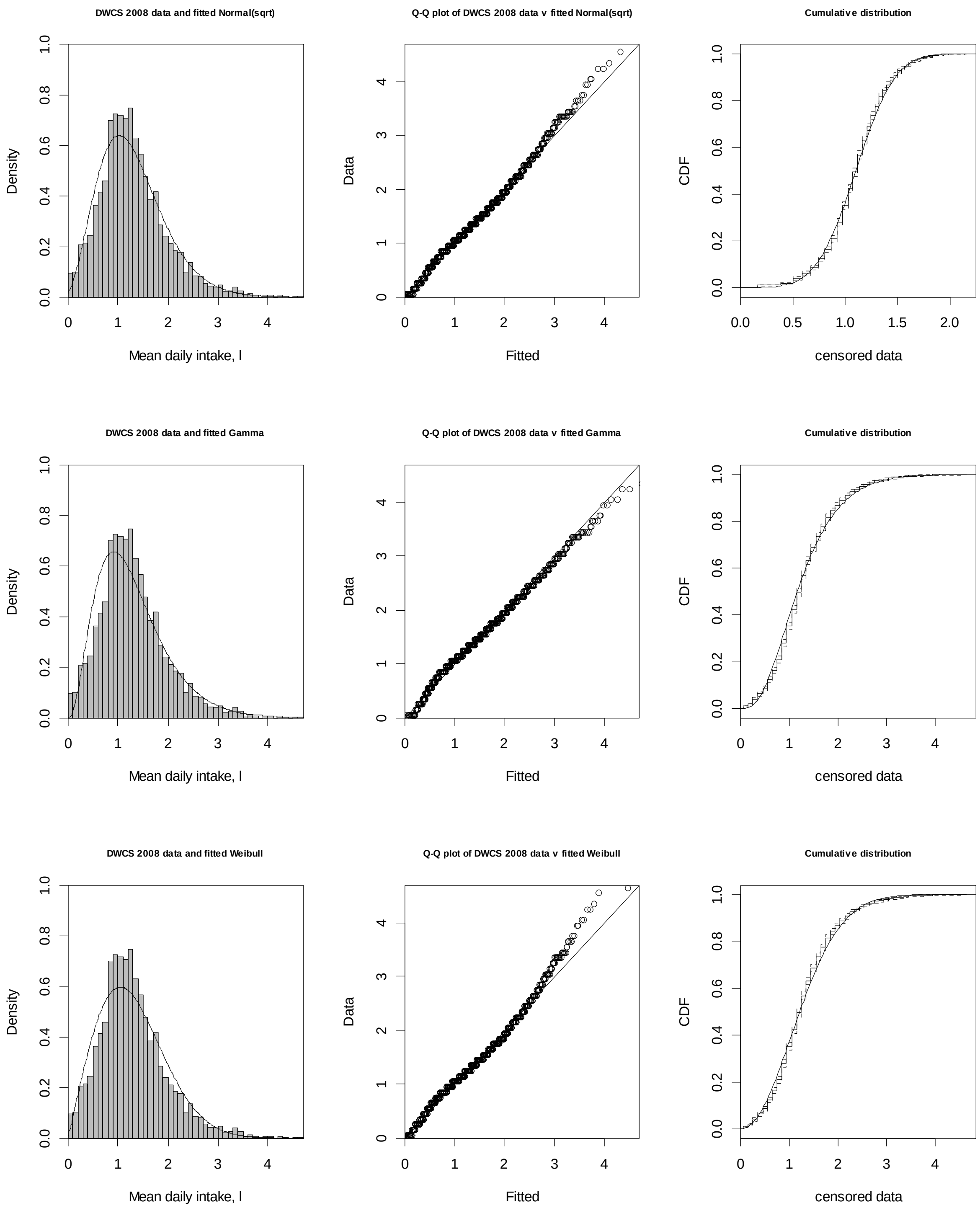


Figure 9. Normal (square root), gamma and Weibull distributions fitted to total tap water intake (Wave 1 + Wave 2) from DWCS 2008 using interval method (error bounds on the empirical CDF computed by *fitdistcens* in R). Note that the CDF for the normal (square root) distribution is in transformed units.

3.2 DWCS 1995

As discussed in Section 2.3 the data for DWCS 1995 were derived from the histogram in the report, so the second method of fitting described above was applied. The normal (square root) distribution fitted slightly better than the others according to the log likelihood and the Weibull distribution fitted better than the gamma distribution (Table 12). The Weibull gave the closest agreement with the 99th percentile, and the gamma distribution gave the worst. However, the Q-Q plot showed that both the normal and the Weibull systematically underestimated the quantiles beyond the 99th percentile, while the errors in the gamma distribution in this range were generally smaller.

Table 12. Log likelihood and 99th percentile for distributions fitted to total water intake from DWCS 1995 using *fitdistcens*.

Distribution	Log likelihood	99th percentile of data	99th percentile of distribution
Normal (sqrt)	-2314	2.875	2.965
Weibull	-2316		2.846
Gamma	-2334		3.200

3.3 DWCS 1978

The same method was applied to the data from DWCS 1978. In this case, the gamma distribution fitted best by likelihood, estimation of the 99th percentile and inspection of the Q-Q plot (Table 13). The normal (square root) distribution fitted the 99th percentile more closely than the Weibull distribution.

Table 13. Log likelihood and 99th percentile for distributions fitted to total water intake from DWCS 1978 using *fitdistcens*.

Distribution	Log likelihood	99th percentile of data	99th percentile of distribution
Normal (sqrt)	-11614	2.475	2.226
Weibull	-11677		2.171
Gamma	-11591		2.346

3.4 NDNS 2001

The NDNS 2001 data were supplied as detailed intake records, from which the daily average intake for each individual was calculated. It was, therefore, possible to fit the distributions directly to the data and to calculate the RSS as a measure of the distance between the empirical and fitted cumulative distribution functions. The models were fitted to the whole sample and also to the subsets by age and sex described in Sections

2.7 and 2.8. The corrected value of the log likelihood for the normal (square root) distribution was used. The Weibull distribution consistently gave the poorest fit as measured by the RSS and, in all cases but one, as measured by the likelihood and the 99th percentile, so it was rejected. There was usually little visible difference between the distributions, thought the Weibull distribution generally had a flatter peak than the other two (as in the example shown Figure 10).

Table 14. Log likelihood, 99th percentile and RSS of the CDF for distributions fitted to total tap water intake from NDNS 2001.

Data set	Distribution	Corrected log-likelihood	99th percentile of data	99th percentile of distribution	RSS of CDF
All	Normal (sqrt)	-1350	2.855	2.710	0.0221
	Gamma	-1354		2.902	0.0269
Females	Normal (sqrt)	-735	2.721	2.661	0.0189
	Gamma	-737		2.857	0.0227
Males	Normal (sqrt)	-611	2.972	2.760	0.0578
	Gamma	-612		2.947	0.0484
Age 19–24	Normal (sqrt)	-70	2.168	2.021	0.0438
	Gamma	-68		2.167	0.0715
Age 25–39	Normal (sqrt)	-484	2.863	2.626	0.0138
	Gamma	-486		2.822	0.0326
Age 40–54	Normal (sqrt)	-492	2.916	2.804	0.0450
	Gamma	-494		2.986	0.0486
Age 55–64	Normal (sqrt)	-257	2.848	2.719	0.0577
	Gamma	-255		2.863	0.0647

The likelihoods for the two fitted distributions were similar in all cases (Table 14). The gamma distribution usually gave the best estimate of the 99th percentile but usually higher RSS. The CDF and the Q–Q plot generally showed very little difference between the two distributions and both fitted the data well over most of the range. Where there was a difference, the Q–Q plot showed that the gamma distribution fitted better at the highest intakes.

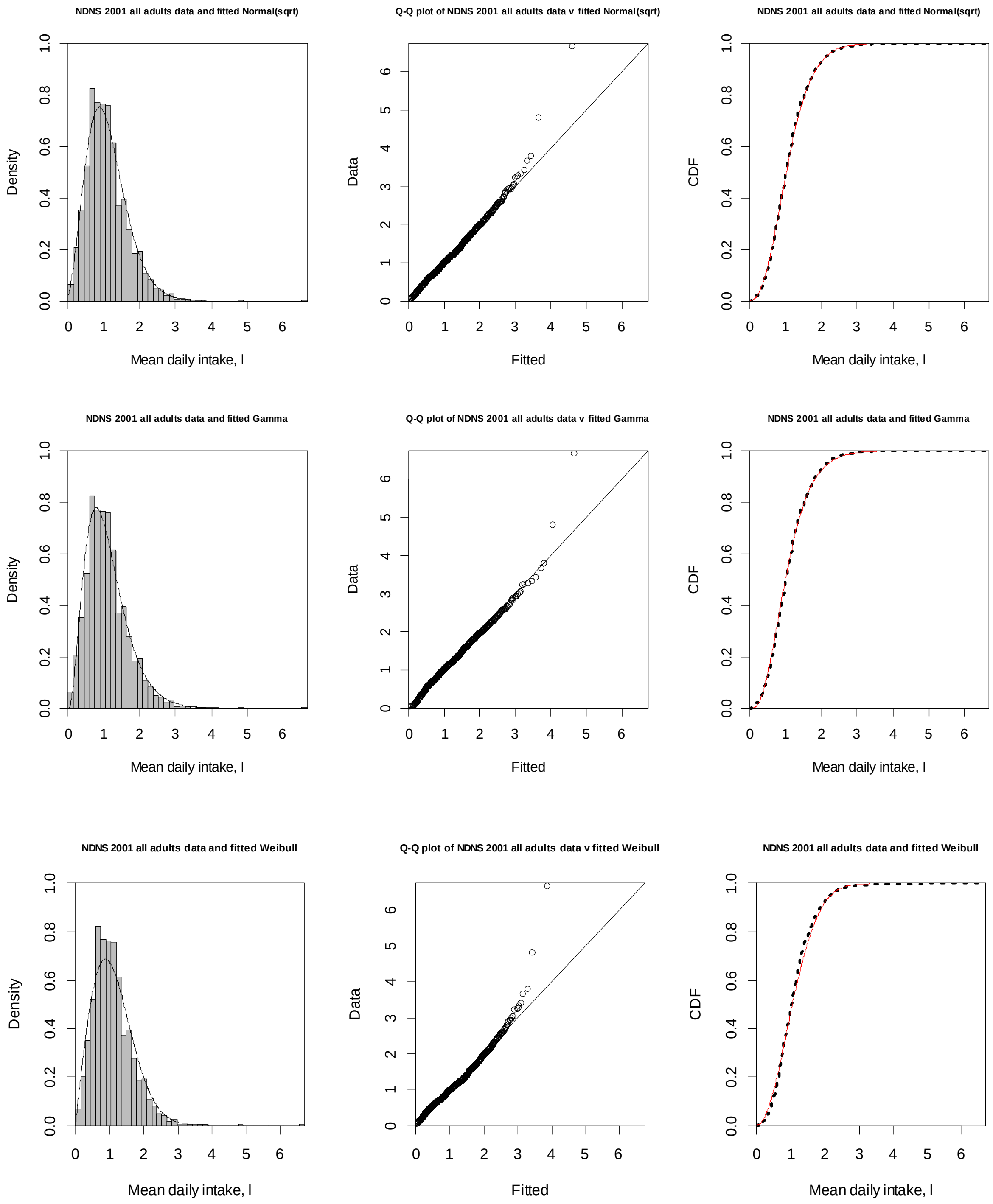


Figure 10. Normal (square root), gamma and Weibull distributions fitted to total tap water intake by all adults from NDNS 2001.

3.5 DW16 2011

The normal (square root) and gamma distributions were fitted to the data for the complete sample in DW16 2011 and to the groups defined by the sex, age and weight factors using the methods that were applied to the NDNS 2001 data set (Table 15). As with the other data sets, the differences between the two distributions were small.

Table 15. Log likelihood, 99th percentile and RSS of the CDF for distributions fitted to tap water intake from DW16 2011.

Group	Distribution	Log-likelihood	99th %ile of data	99th %ile of distribution	RSS of CDF
All	Normal (sqrt)	-329	1.704	1.597	0.0167
	Gamma	-332		1.783	0.0123
Females	Normal (sqrt)	-116	1.694	1.477	0.0179
	Gamma	-124		1.661	0.0439
Males	Normal (sqrt)	-204	1.705	1.709	0.0163
	Gamma	-201		1.893	0.0149
Age 0–3	Normal (sqrt)	-30	1.466	1.366	0.0246
	Gamma	-32		1.558	0.0565
Age 4–6	Normal (sqrt)	-0.58	1.177	1.224	0.0206
	Gamma	-1.20		1.339	0.0287
Age 7–10	Normal (sqrt)	-93	1.700	1.629	0.0333
	Gamma	-94		1.799	0.0281
Age 11–15	Normal (sqrt)	-153	1.910	1.881	0.0140
	Gamma	-156		2.091	0.0540
< 17 kg	Normal (sqrt)	-21	1.527	1.345	0.0362
	Gamma	-21		1.530	0.0526
17–26 kg	Normal (sqrt)	-18	1.194	1.317	0.0401
	Gamma	-18		1.429	0.0215
26–41 kg	Normal (sqrt)	-65	1.715	1.625	0.0209
	Gamma	-72		1.858	0.0925
≥ 41 kg	Normal (sqrt)	-122	1.960	1.968	0.0293
	Gamma	-126		2.185	0.1073

3.6 Choice of distribution for simulation

The three distributions that fitted the total tap water intake best were Weibull, gamma and normal fitted to the square root of the intake. Overall, the third of these gave the best fit, but the gamma distribution had the least tendency to underestimate the upper quantiles. There is no strong theoretical justification for choosing any one of the distributions. The Weibull and the gamma typically arise in waiting time or failure interval distributions. The use of the normal distribution for the square root of intake was motivated by the observation that this approximates a Poisson distribution. However, the assumption of constant variance was relaxed, so this was not in fact the normal approximation to the Poisson distribution. Furthermore, the Poisson distribution is appropriate for the frequency of independent events of equal probability, so there is little theoretical basis for using it as a model of individual intake, where the long-term mean intakes may differ between individuals.

The choice of distribution was, therefore, largely pragmatic. It was convenient, though not essential to use one distribution and vary the parameters to represent every data set. On this basis the Weibull distribution was the least suitable, as it gave poor results with NDNS 2001.

In practice, simulation of exposure using either the normal (square root) or the gamma distributions would be similar: the Q-Q plots showed that there was very little difference in their estimation of the quantiles over most of the range. In general, the normal (square root) tended to underestimate the highest quantiles more often than the gamma distribution. This means that, compared with the normal distribution, the fitted gamma distribution would tend to predict higher exposure estimates for the same (high percentile) probability, which is more conservative (precautionary). The other cause for doubt about the use of the normal distribution is the need to truncate the underlying distribution at zero. This would be necessary because it was fitted to the positive square root of the data. The positive section of the distribution would then be normalised to give a total probability density of 1.

It was concluded that using the gamma distribution in the simulation was preferable. The parameters for all the data sets are given in Table 16.

Table 16. Parameters (shape and rate) of the fitted gamma distributions for all the tap water intake data sets.

Data set	Shape	Rate
DWCS 2008 total	3.44	2.66
DWCS 2008 Wave 1	3.64	2.85
DWCS 2008 Wave 2	3.25	2.47
DWCS 1995	3.03	2.64
DWCS 1978	4.24	4.44
NDNS 2001 all	3.53	3.20
NDNS 2001 males	3.72	3.26
NDNS 2001 females	3.41	3.18
NDNS age 19–24	2.92	3.81
NDNS age 25–39	3.34	3.18
NDNS age 40–54	3.99	3.36
NDNS age 55–64	4.49	3.78
DW16 2001 All	2.08	3.81
DW16 2001 Females	2.08	4.08
DW16 2001 Males	2.13	3.63
DW16 2001 Age 0–3	1.76	3.97
DW16 2001 Age 4–6	2.61	5.78
DW16 2001 Age 7–10	2.40	4.10
DW16 2001 Age 11–15	2.28	3.42
DW16 2001 Weight < 17 kg	1.71	3.98
DW16 2001 Weight 17–26 kg	2.81	5.65
DW16 2001 Weight 26–41 kg	2.03	3.60
DW16 2001 Weight ≥ 41 kg	2.37	3.35

4. Concentration of chemicals in tap water

Sets of data on measured concentrations of chemical contaminants in water were supplied by the DWI. These consisted of the data collected by all the water companies in England and Wales for statutory monitoring of water quality and were supplied with their permission for use in the project.

"More than one-third of the tests were carried out on drinking water drawn from consumers' taps selected at random. For monitoring purposes, company water supply areas are divided into zones based on population (maximum 100,000). Generally, zones are sampled at consumers' taps with the number of required tests being greatest in zones with larger populations. Other sample locations are water treatment works and treated water (service) reservoirs" (DWI, 2011). The proportion taken from consumers' taps, as reported in other reports in the same set, was similar in the other regions of England and Wales. With the exception of sulphate and chloride, all the parameters considered in this study (listed in Table 17) must be monitored at customers' taps. Most samples are taken after flushing, to determine the quality at the point of supply. However, samples for lead and copper are the first litre drawn during a random daytime visit. Consequently samples for lead and copper will reflect any increases due to domestic plumbing. Concentrations of lead and copper in flushed samples will almost certainly be lower than in regulatory samples.

Given the large number of samples and the randomisation employed, it is assumed here that the results are representative of supplies to consumers in England and Wales as a whole. There are some local variations, but these are generally smaller than the variations in the distribution of nitrate shown in Appendix C.

The main sets contained the concentrations of ten inorganic determinands for 2010 (Table 17), the concentrations of iron, lead selenium, sodium and manganese for 2004, and of lead for the intervening years. In addition, frequency tables for lead concentration from 1994–96 were taken from printed reports by optical character recognition (Hydes *et al.*, 1996; Hydes *et al.*, 1997). From 2004 onwards, each record gave the value and a flag to indicate whether this was the measured concentration or the limit of detection (LoD), in which case the concentration was an unknown value up to the LoD. For all the determinands listed in Table 17 the LoD must be no more than one-tenth of the limit for the concentration prescribed in the Water Supply Regulations (United Kingdom, 2000).

Table 17. Determinands present in the data set for 2010.

Determinand	Prescribed concentration	Maximum permitted LoD	Units	Number of samples	Number < LoD
Sulphate*	250.0	25.0	mg	10,524	48
Sodium	200.0	20.0	mg	18,234	7
Nitrate	50.0	5.0	mg	24,378	0
Iron	200.0	20.0	µg	45,684	17,376
Copper	2.0	0.2	mg	13,029	1,322
Arsenic	10.0	1.0	µg	12,825	3,958
Lead	25.0	2.5	µg	12,667	6,323
Selenium	10.0	1.0	µg	12,646	5,428
Chloride*	250.0	25.0	mg	10,536	2
Manganese	50.0	5.0	µg	41,420	24,775

* Indicators

Initial exploration concentrated on the data for 2010. The project specified that lead, iron and selenium should be modelled, together with one additional determinand to be agreed during the project. Manganese was subsequently added. After examining the data, it was proposed that the additional substance modelled should be one with different statistical properties from the first three. Iron and lead concentration both had highly skewed distributions with the majority of samples very close to zero, but long tails containing sparse results at much higher concentrations. Both had high proportions below the LoD. The distribution of selenium concentration was less skewed, but still had a high proportion below the LoD.

The distribution for copper had a similar shape to those for lead and iron, but many fewer points below the LoD, so it was used to inform the analyses of lead and iron, but not selected for full consideration. Arsenic had a similar distribution to selenium, with a similar proportion below the LoD, so was not considered further. Sodium, chloride and sulphate had less skewed distributions with very few values below the LoD. Sodium was, therefore, chosen as a representative of this group.

The distribution for nitrate was complicated. At national level it was strongly bimodal, with peaks around 2 mg/l and 30 mg/l. When the data were viewed at company level, many still showed multimodal behaviour, confirming our expectation that exposure will vary with region and water source within each region. Some sources, such as those from upland catchments, could have consistently low concentrations. Others, such as surface waters from draining catchments with a high fraction of arable land, may have high seasonal peak concentrations. Finally, some groundwater sources have persistently raised levels of nitrate, particularly those where land has been converted from permanent grassland to arable farming in the past (which releases a lot of nitrate via enhanced mineralisation) and where land is intensively managed (including fertilised grassland). The type of analysis required to interpret these drivers was beyond the scope of this project. Some examples of the distributions are shown in Appendix C.

The large number of samples with concentrations below the LoD for some of the determinands was one of the main concerns when considering how to use this data in the simulation of exposure. Although they would have little influence on the upper quantiles, which were likely to be of most interest in the assessment, they amounted to over half the sample in the case of lead and so would affect the median value. The LoD was not a constant for each determinand: its value varied between and within the water companies. In some cases, concentrations were recorded for some samples that were below the LoD for others. This meant that points recorded as below the LoD could affect quantiles higher than those suggested by their frequency alone.

As noted in Section 1.2.1, several substitution methods are commonly used. The one recommended by the FSA is to repeat the simulation using two substitutions: 0 and equal to the LoD. Another substitution approach is to use LoD/2. This problem has been examined in detail by a working group of the European Food Safety Authority (EFSA, 2010). They commented that

> Despite its drawbacks, the substitution method is still widely used, mainly with the justification that it is easy to implement, it is widely understood and that the upper bound practice leads to conservative estimates for exposure assessment calculations, i.e. over-estimation of the mean and under-estimation of the variability.

Clearly substitution by 0 can lead to the converse: under-estimation of the mean and over-estimation of the variability.

The EFSA working group considered several parametric methods for fitting distributions while taking account of points below the LoD and commented:

> The consensus view is that the maximum likelihood estimation (MLE) method is the best approach from a methodological perspective.

They recommend against proceeding with any method if there are than fewer than 50 samples or more than 80% are censored, neither of which applies in this case. Otherwise parametric modelling is generally the preferred method, with the option of the Kaplan-Meier method (Kaplan & Meier, 1958) if fewer than 50% of the data are censored and there are multiple LoDs. Kaplan-Meier is a non-parametric method that essentially constructs empirical distribution functions with uncertainty arising from the multiple LoDs. It is probably better suited to estimating population statistics than to use in simulation.

A recent paper dealing with the same issue in environmental monitoring (Gardner, 2011) reviewed several earlier papers and constructed examples to test substitution and MLE approaches. It concluded:

- Perhaps the only justifiable substitution methodology is where double substitution (by zero and then by the LOD value) is used to give an upper and lower bound to a mean value. [The method used by the FSA]

- Use of MLE is recommended as for providing a better approach and should be seen as essential if there is a need to estimate standard deviation or percentile values.

On the basis of these and other papers, we proceeded with the use of MLE to attempt to fit distributions to the contaminant data. The method used, as with the interval data for intake, employed the *fitdistcens* function from the *fitdistrplus* toolbox for R. The censored values are specified as intervals in which the upper bound is the LoD and the lower bound is NA (signifying a missing value). In contrast to the MLE methods used in the EFSA (2010), this allows multiple LoD values within the data set.

EFSA (2010) also considered goodness of fit statistics. The use of the Akaike and Bayes information criteria (AIC and BIC) to discriminate between models was suggested. However, as observed above, these are equivalent to using the likelihood when all the models have the same number of parameters. The report contains only a brief paragraph on formal goodness of fit tests. As was the case with the intake distributions (Section 3.1), these did not prove useful in practice.

In addition to the graphical assessment of the fitted distributions, particularly the Q–Q plot, the RSS of the cumulative distribution function was calculated, as was the Kullback-Leibler divergence (Kullback & Leibler, 1951), which is a measure of the distance between two density functions – in this case the empirical and fitted densities.

The data for iron and lead contained a few extreme values: the maximum concentration for each was over 100 times the 99th percentile. After discussion with the DWI it was agreed that these were likely to be genuine outliners, not errors in the analysis or recording, so they were retained in the data sets. Nine values for iron were given as less than LoDs which were much greater than required by the regulations. These were found to be the results of the wrong analysis being used and they were removed from the data set.

4.1 Copper

It was noted that copper concentration had a similar distribution to lead and iron, but with only 10% below theLoD, compared with 40% for iron and 50% for lead. It therefore provided a useful test for suitable distributions. The distribution of the data is shown in Figure 11 up to the 99th percentile (including the remainder of the tail would compress the scale without conveying additional information). Some summary statistics are given in Table 18 showing the results of substitution of values below the LoD by 0 and by the LoD.

Although the shape of the histogram would suggest an exponential distribution, this gave a very poor fit. The Q–Q plot was a curve lying below the 1:1 line (i.e. overestimating) up to 0.09 mg/l (the 90th percentile) and above the line (underestimating) beyond that point. The comparison of the density function with the histogram and between the CDFs also showed the poor fit very clearly.

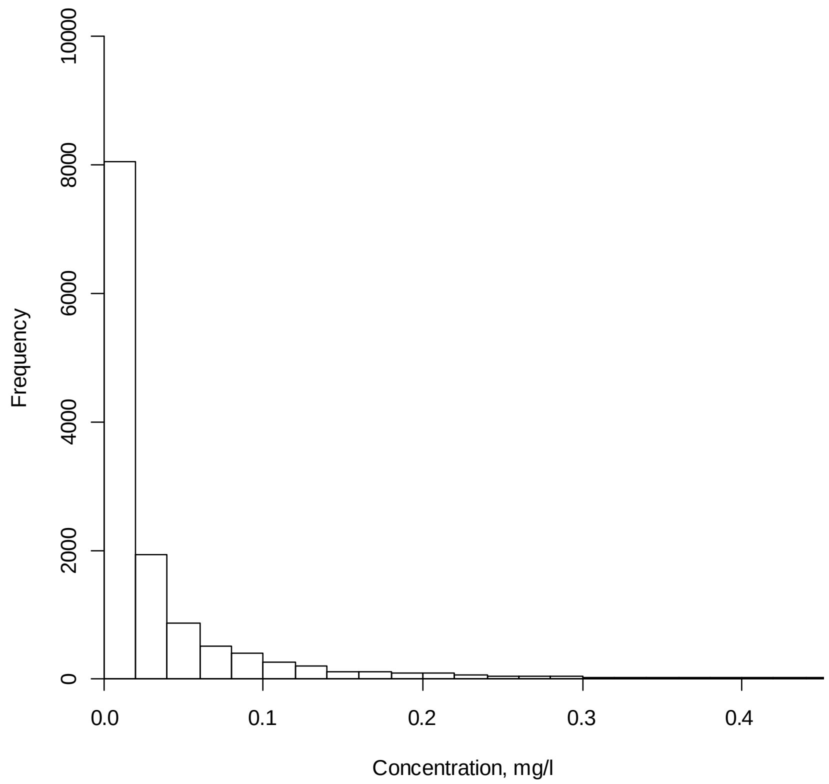


Figure 11. Histogram of copper concentration in all samples from England and Wales 2010 (up to 99th percentile).

Table 18. Summary statistics for copper concentration in all samples from England and Wales 2010 using two substitutions for values below the LoD.

Statistic	Value, mg/l <LoD = 0	Value, mg/l <LoD = LoD
Mean	0.0415	0.0419
Minimum	0.0000	0.0001
Maximum	2.7000	2.7000
Percentile 25	0.0050	0.0053
Percentile 50	0.0130	0.0130
Percentile 75	0.0374	0.0374
Percentile 90	0.0990	0.0990
Percentile 95	0.1700	0.1700
Percentile 99	0.4470	0.4470
Percentile 99.9	1.0401	1.0401

The only standard parametric distribution that fitted the data was the lognormal (parameters $\mu = -4.282$, $\sigma = 1.496$), which gave a very good agreement up to the 90th percentile, but overestimated beyond that point, though the difference was small even at the 99th percentile (Table 19, Figure 12). This result was obtained using the MLE method for left-censored data. Substitution of LoD values by 0 before fitting the distribution resulted in underestimation of the lower percentiles and severe overestimation from the 70th percentile onwards. Conversely, substitution by the LoD caused underestimation above the 75th percentile and substituting LoD/2 produced similar, but less severe underestimation.

Table 19. Summary statistics from a lognormal distribution fitted to data for copper from England and Wales.

Statistic	Value, mg/l
Mean	0.0423
Min	0.0
Max	∞
Percentile 25	0.0050
Percentile 50	0.0138
Percentile 75	0.0379
Percentile 90	0.0940
Percentile 95	0.1618
Percentile 99	0.4485
Percentile 99.9	1.4063

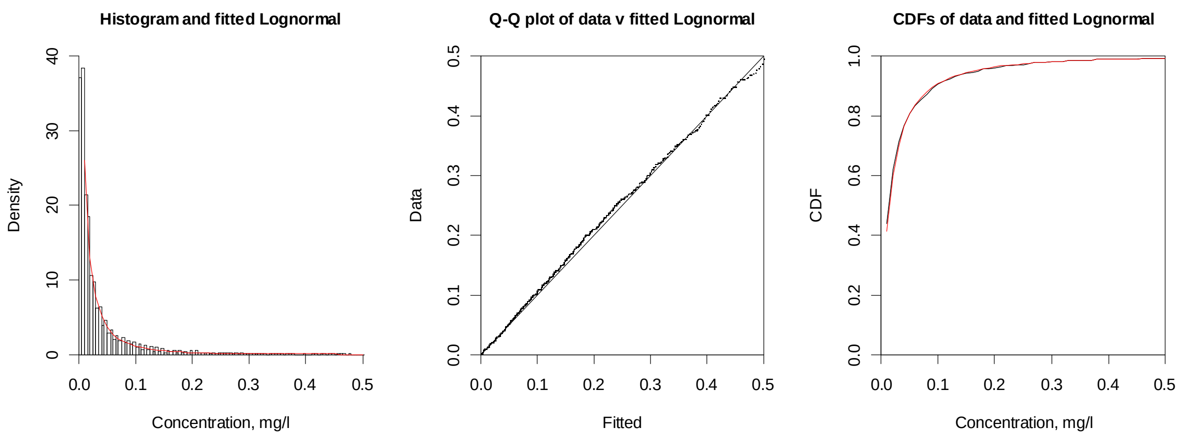


Figure 12. Comparison of fitted lognormal with concentration data for copper from England and Wales 2010 up to 99th percentile.

The Kullback-Leibler (KL) divergence proved to be of little help in selecting the distribution to use. Despite giving a very poor fit, the KL divergence for the exponential distribution was 0.918 compared with 1.012 for the lognormal distribution. The log likelihood for the lognormal distribution was higher (24312 compared with 20709), giving a corresponding difference in the AIC and BIC. The RSS computed for the CDF also showed a much smaller error for the lognormal (0.0002 compared with 0.057).

Using the lognormal distribution appeared to give a satisfactory fit to the data, except in the tail of the distribution and, in particular, to provide a better method of handling values below the LoD in the simulation than fixed substitutions. However, the tail is important for exposure assessment, and the use of fitted distributions will be discussed further in Section 6. The methods that worked best for copper were applied to the other contaminants.

4.2 Lead

The permitted concentration of lead in tap water was reduced from 50 µg/l to 25 µg/l on 25th December 2003 and is due to be reduced to 10 µg/l on 25th December 2013. The data showed how the actual concentration changed over the period 1994–2010 (Table 20, Table 21 and Figure 13). The distribution functions in the figure are lognormal distributions fitted to the data sets as described below. The distribution slightly underestimated the frequency of very rare high values, but these lie beyond the range and the resolution of the graph. The substitution used for the LoD in Table 21 had a very large effect on the median, due to the high proportion of values below the LoD, and a substantial effect on the mean.

These results show that the concentration of lead has been reduced since 1994 and that the chance of exceeding the maximum acceptable concentration has substantially decreased. There also seems to have been a reduction between 2004–07 and 2008–10, probably as a result of the introduction of plumbosolvency schemes in order the meet the tightening standard. It should be noted that all samples are taken in consumers' homes and most of the lead enters the water from lead supply pipes, much of which is owned by the consumer and was probably installed before 1970. Some lead may arise from use of non-approved solder and leaching from brass fittings.

Table 20. Proportion of samples where lead concentration was below past, current and future limits.

Year	10 µg/l	25 µg/l	50 µg/l
1994	0.812	0.912*	0.969
1995	0.781	0.896*	0.961
1996	0.837	0.934*	0.977
2004	0.9722	0.9950	0.9984
2005	0.9804	0.9970	0.9991
2006	0.9780	0.9966	0.9989
2007	0.9824	0.9976	0.9993
2008	0.9887	0.9977	0.9991
2009	0.9900	0.9977	0.9993
2010	0.9893	0.9983	0.9992

* This value is not available in the reports for 1994–96, so is interpolated between the proportions for 20 and 30. It is likely to be a slight underestimate.

Table 21. Summary statistics for lead concentration in all samples from England and Wales 2004, 2010 using two substitutions for values below the LoD.

Year	2004	2004	2010	2010
Substitution	<LoD = 0	<LoD = LoD	<LoD = 0	<LoD = LoD
Statistic	Value, µg/l	Value, µg/l	Value, µg/l	Value, µg/l
Mean	1.79	2.04	1.01	1.21
Minimum	0.0	0.00	0.00	0.02
Maximum	1772.20	1772.20	1530.00	1530.00
Percentile 25	0.00	0.50	0.00	0.25
Percentile 50	0.20	0.70	0.02	0.50
Percentile 75	1.53	1.53	0.60	0.80
Percentile 90	4.20	4.20	2.00	2.00
Percentile 95	6.90	6.90	3.77	3.77
Percentile 99	17.51	17.51	10.07	10.07
Percentile 99.9	88.51	88.51	40.31	40.31

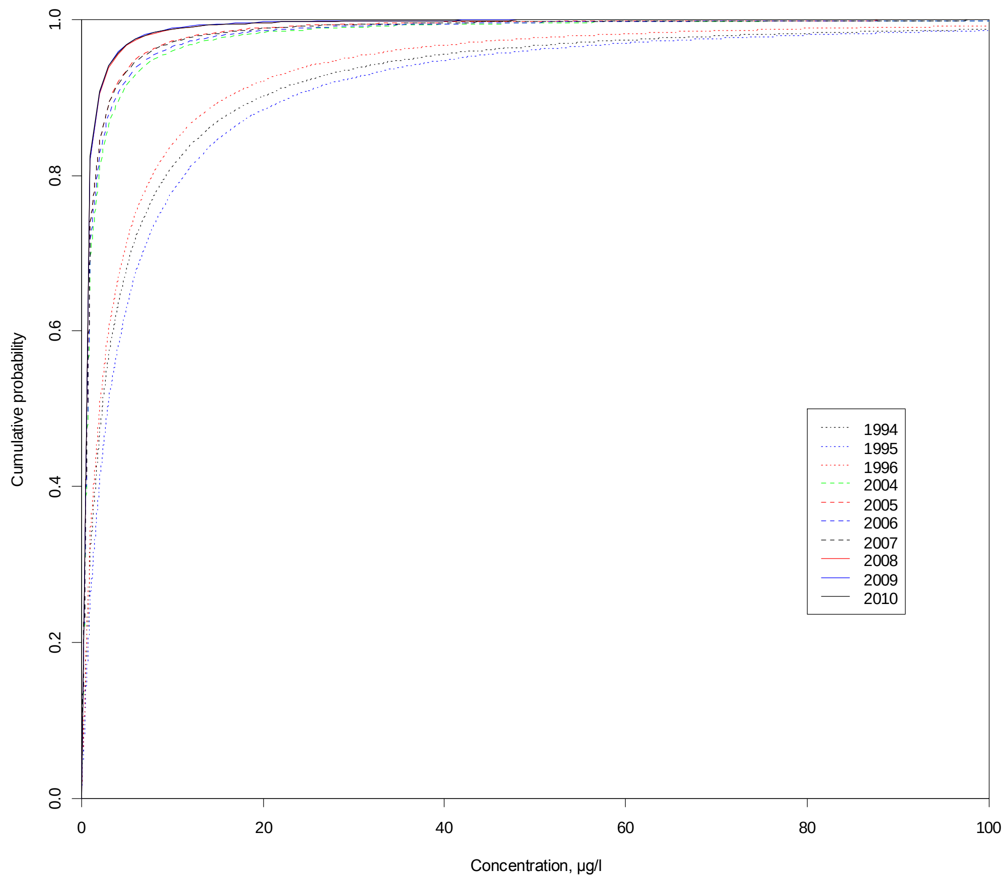


Figure 13. Change over time in the cumulative distribution functions for the concentration of lead in tap water (using lognormal distributions fitted to the data).

Distributions were fitted to the data sets containing complete sets of sample results (2004–2010) using the MLE method for left-censored data. The exponential distribution was tested for each set, but a better fit was obtained consistently using the lognormal distribution. As was the case for copper, some divergence was seen in the tail of the distribution. For 2008–2010 this began between the 95th and 99th percentiles; for 2004–2007 it began between the 90th and 95th percentiles.

As the data for 1994–96 were only available as class frequencies, lognormal distributions were fitted by applying the method used for the DWCS intake surveys. The observed and predicted class frequencies were compared to verify that the results were acceptable. The parameters are shown in Table 22. It can be seen that μ, which is directly related to the geometric mean, had decreased considerably over time. The summary statistics are shown in Table 23.

Table 22 Parameters for the lognormal distributions fitted to lead concentrations.

Year	μ	σ
1994	0.807	1.698
1995	1.06	1.62
1996	0.679	1.636
2004	−0.946	1.861
2005	−1.129	1.796
2006	−1.062	1.862
2007	−1.161	1.84
2008	−1.59	1.746
2009	−1.535	1.69
2010	−1.636	1.76

Table 23. Summary statistics for lognormal distributions fitted to lead concentration data from England and Wales (μg/l).

Year	1994	2004	2010
Mean	9.47	2.19	0.92
Min		0.0	0.0
Max	∞	∞	∞
Percentile 25	0.71	0.11	0.06
Percentile 50	2.24	0.39	0.19
Percentile 75	7.04	1.36	0.64
Percentile 90	19.75	4.22	1.86
Percentile 95	36.60	8.29	3.52
Percentile 99	116.41	29.47	11.69
Percentile 99.9	425.90	122.11	44.83

4.3 Iron

The data for iron showed some decrease in concentration between 2004 and 2010 (Table 24). The 99th percentile (1 in 100) is now about half of the limit of 200 μg/l prescribed by the Water Supply Regulations, although the 99.9th percentile (1 in 1000) is about double the limit. The summary statistics are shown using two substitutions for values below the LoD: 0 and the LoD. As expected, with 40% of samples below the LoD, the substitution had the biggest effect on the lower quartile and the median; the mean changed by about 25% and the upper percentiles were unaffected.

The distributions were fitted to the data for iron from 2004 and 2010 using the MLE method for left-censored data. The lognormal distribution gave good results for 2010 up to the 99th percentile, after which it underestimated significantly. A similar pattern was found with the data for 2004, but the fit was poorer throughout the range. The parameters were $\mu = 2.281$, $\sigma = 1.257$ for 2004 and $\mu = 1.729$, $\sigma = 1.32$ for 2010, again reflecting the decrease in the mean concentration. The resulting summary statistics are shown in Table 25 for comparison with the data. As expected, the mean and median lie between the estimates produced using the two methods of LoD substitution.

Table 24. Summary statistics for iron concentration in all samples from England and Wales 2004, 2010 using two substitutions for values below the LoD.

Year	2004	2004	2010	2010
Substitution	<LoD = 0	<LoD = LoD	<LoD = 0	<LoD = LoD
Statistic	Value, µg/l	Value, µg/l	Value, µg/l	Value, µg/l
Mean	18.3172	22.7215	12.2258	15.1596
Minimum	0.0	0.007	0.0	1.0
Maximum	4586.0	4586.0	2260.0	2260.0
Percentile 25	0.0	10.0	0.0	4.8
Percentile 50	10.0	13.0	4.3	10.0
Percentile 75	21.5	21.5	14.0	15.0
Percentile 90	42.1	42.1	28.7	28.7
Percentile 95	64.0	64.0	45.0	45.0
Percentile 99	142.0	142.0	116.0	116.0
Percentile 99.9	511.2	511.2	402.8	402.8

Table 25. Summary statistics for lognormal distributions fitted to iron concentration data from England and Wales (µg/l).

Year	**2004**	**2010**
Mean	21.6	13.5
Min	0.0	0.0
Max	∞	∞
Percentile 25	4.2	2.3
Percentile 50	9.8	5.6
Percentile 75	22.8	13.7
Percentile 90	49.0	30.6
Percentile 95	77.4	49.4
Percentile 99	182.2	121.5
Percentile 99.9	476.0	333.0

4.4 Selenium

The data for selenium showed some decrease in concentration between 2004 and 2010 (Table 26). The 99.9th percentile (1 in 1000) is now less than half of the limit of 10 µg/l prescribed by the Water Supply Regulations.

Table 26. Summary statistics for selenium concentration in all samples from England and Wales 2004, 2010 using two substitutions for values below the LoD.

Year Substitution Statistic	2004 <LoD = 0 Value, µg/l	2004 <LoD = LoD Value, µg/l	2010 <LoD = 0 Value, µg/l	2010 <LoD = LoD Value, µg/l
Mean	0.71	0.85	0.41	0.57
Minimum	0.0	0.0004	0.0	0.06
Maximum	8.80	8.80	7.40	7.40
Percentile 25	0.0	0.30	0.00	0.24
Percentile 50	0.45	0.56	0.30	0.40
Percentile 75	1.00	1.00	0.60	0.80
Percentile 90	1.90	1.90	1.00	1.00
Percentile 95	2.50	2.50	1.30	1.30
Percentile 99	4.15	4.15	2.64	2.64
Percentile 99.9	6.48	6.48	4.58	4.58

The distribution was less skewed than those for copper, lead and iron (Figure 14). In both 2004 and 2010, over 30% of the values were below the LoD, which was often at or close to maximum permitted (1 µg/l).

The candidate distributions were fitted using the MLE method for left-censored data. Only the exponential distribution gave acceptable results for the 2004 data. The lognormal distribution fitted this set very poorly: the RSS of the CDF was 0.85 compared with 0.06 for the exponential distribution. The lognormal distribution overestimated severely (by a factor of up to 4) from the 80th percentile upward. The choice of the exponential distribution would not have been obvious from initial examination of the data, because the exponential distribution has its mode at zero, whereas the data appears to have a non-zero mode. However, the data set contains a large proportion of values below the LoD: the three peaks in Figure 14 and many of the lower bars at the left of the histogram correspond to LoD values from different sets of samples. These therefore represent true values lying between 0 and the LoD, so the mode cannot be estimated reliably from the data. The exponential distribution fitted well in the middle region, but underestimated slightly beyond the 95th percentile. Its single parameter, the rate, was 1.297.

The data for 2010 showed some evidence of containing two populations with different means. The lognormal distribution and the exponential distribution both gave similar goodness of fit results with this set. For example the RSS of the CDF was 0.15 for the lognormal distribution and 0.18 for the exponential distribution. The lognormal distribution overestimated slightly at the higher percentiles, whereas the exponential distribution underestimated slightly. To be consistent with 2004, the exponential distribution was selected; its rate parameter was 2.111. The summary statistics for both years are shown in Table 27. The mean, median and lower quartile (and upper quartile for 2010) fall in the ranges shown for the two methods of LoD substitution in Table 26.

The difference in distribution between selenium and the metals considered above may be due to their origins. Selenium in tap water arises mainly from naturally-occurring selenium in the water sources, whereas most copper, and lead enter the water from the supply pipes, including domestic pipework and fittings, and iron can arise from treatment or corrosion of the iron distribution mains.

Table 27. Summary statistics for exponential distributions fitted to selenium concentration data from England and Wales (μg/l).

Statistic	2004	2010
Mean	0.77	0.47
Min	0.0	0.0
Max	∞	∞
Percentile 25	0.22	0.14
Percentile 50	0.53	0.33
Percentile 75	1.07	0.66
Percentile 90	1.78	1.09
Percentile 95	2.31	1.42
Percentile 99	3.55	2.18
Percentile 99.9	5.33	3.27

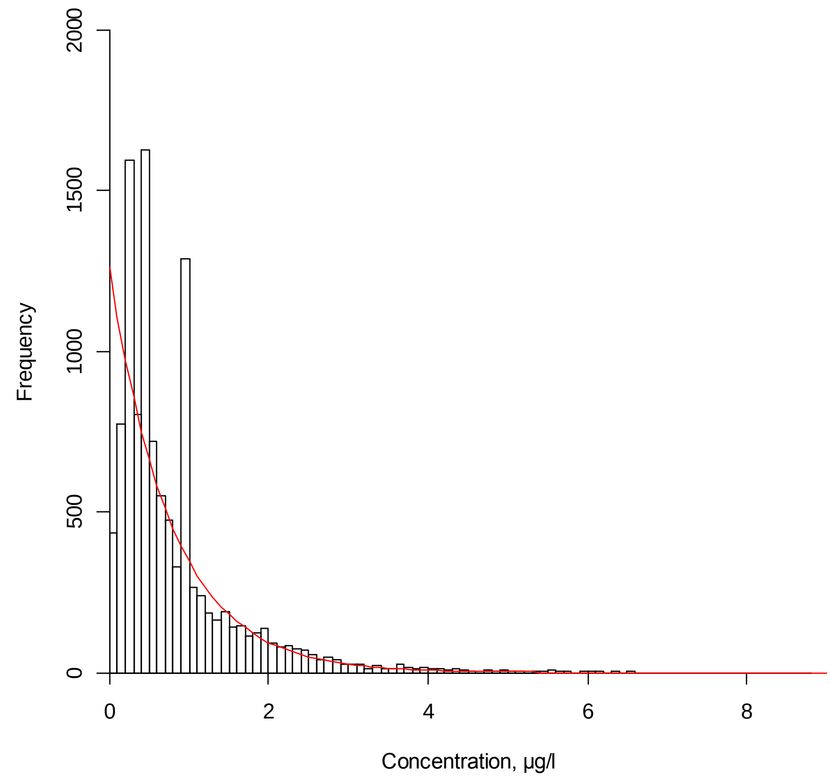


Figure 14. Histogram of selenium concentration in all samples from England and Wales 2004 with fitted exponential distribution.

4.5 Sodium

Sodium is present in some sources of ground water and surface. Some treatment chemicals contain sodium, and significant quantities may be added by domestic water softeners. The data for sodium showed a slight increase in concentration between 2004 and 2010 (Table 28), but the 99.9th percentile (1 in 1000) is still less than the limit of 200 mg/l prescribed by the Water Supply Regulations.

The distribution was less skewed than those for copper, lead and iron, with a non-zero mode. As there were very few values below the LoD, this appeared to be a genuine mode, so an exponential distribution would be inappropriate. The distributions were fitted to the data using the MLE method for left-censored data. The best fit was, again, obtained with a lognormal distribution, but the Q–Q plot for both years showed systematic errors throughout the range and severe underestimation of the percentiles above the 99th (Figure 15 and Table 29). The parameters of the distribution were $\mu = 2.663$, $\sigma = 0.696$ for 2004 and $\mu = 2.738$, $\sigma = 0.691$ for 2010.

Table 28. Summary statistics for sodium concentration in all samples from England and Wales 2004, 2010.

Year Statistic	2004 Value, mg/l	2010 Value, mg/l
Mean	18.3	19.6
Minimum	0.1	0.1
Maximum	257.0	223.0
Percentile 25	8.4	9.4
Percentile 50	13.6	15.3
Percentile 75	25.0	26.2
Percentile 90	35.8	37.0
Percentile 95	44.0	44.8
Percentile 99	64.8	66.0
Percentile 99.9	150.1	162.5

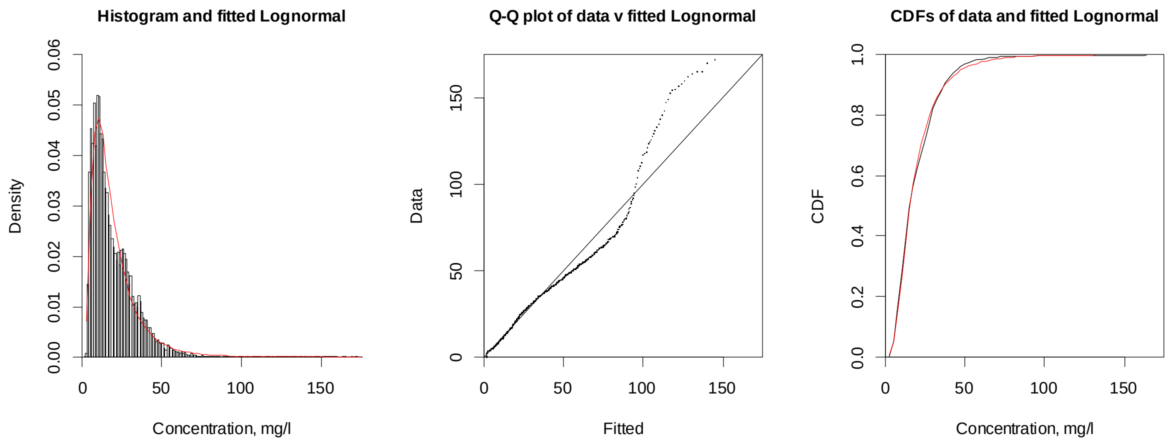


Figure 15. Comparison of fitted lognormal with concentration data for sodium from England and Wales 2010 up to 99.9th percentile.

Table 29. Summary statistics for lognormal distributions fitted to sodium concentration data from England and Wales (mg/l).

Statistic	2004	2010
Mean	18.27	19.62
Min	0.0	
Max	∞	∞
Percentile 25	8.97	9.70
Percentile 50	14.34	15.46
Percentile 75	22.93	24.63
Percentile 90	34.99	37.47
Percentile 95	45.05	48.16
Percentile 99	72.40	77.13
Percentile 99.9	123.20	130.76

4.6 Manganese

The data for manganese showed some decrease in concentration between 2004 and 2010 (Table 30). In 2010, the 99th percentile (1 in 100) was about one-quarter of the limit of 50 µg/l prescribed by the Water Supply Regulations and the 99.9th percentile (1 in 1000) was below the limit, whereas it was slightly above the limit in 2004. In 2004, manganese was detected above the limit of detection in only about 46% of samples; in 2010 this had fallen to 40% although the mean LOD used had been reduced from 2.6 µg/l to 1.7 µg/l.

The summary statistics are shown using two substitutions for values below the LoD: 0 and the LoD. As the proportion of values below the LoD was very high and the range of LoDs used was very wide, including a substantial number at the maximum permitted value of 5.0 µg/l, there were large uncertainties in the lower quartile, median, mean and even the upper quartile. In the data for 2010, this effect extended to the 95th percentile.

Table 30. Summary statistics for manganese concentration in all samples from England and Wales 2004, 2010 using two substitutions for values below the LoD.

Year **Substitution** **Statistic**	2004 <LoD = 0 Value, µg/l	2004 <LoD = LoD Value, µg/l	2010 <LoD = 0 Value, µg/l	2010 <LoD = LoD Value, µg/l
Mean	2.03	3.47	1.10	2.12
Minimum	0.00	0.00	0.00	0.11
Maximum	656.00	656.00	430.00	430.00
Percentile 25	0.00	1.15	0.00	1.00
Percentile 50	0.00	2.64	0.00	1.80
Percentile 75	2.14	4.42	1.20	2.00
Percentile 90	5.52	5.52	2.90	3.39
Percentile 95	8.70	8.70	4.60	5.00
Percentile 99	19.00	19.00	11.90	11.90
Percentile 99.9	65.84	65.84	44.29	44.29

The usual distributions were fitted to the data for manganese from 2004 and 2010 using the methods described above. In general, the distributions fitted quite poorly and were difficult to assess due to the presence of values below the LoD with relatively high LoDs. For 2010, the best fit was obtained using the lognormal distribution with $\mu = -0.286$ and $\sigma = 1.095$. The Weibull distribution appeared to fit the 2004 data slightly better than the lognormal. However, when used to generate random samples, it consistently produced maxima of about 100 µg/l, one-sixth of the maximum in the data set, whereas the lognormal produced maxima around or above that of the data set. We decided to use the lognormal distribution for 2004, to remain consistent with 2010 and with most of the other determinands. The parameters of the lognormal distribution for the 2004 were $\mu = -0.058$ and $\sigma = 1.460$, reflecting the higher mean concentration at that time. The resulting summary statistics from the distributions are shown in Table 31 for comparison with the data. As expected, the mean and median lie between the estimates produced using the two methods of LoD substitution.

Table 31. Summary statistics for lognormal distributions fitted to manganese concentration data from England and Wales (μg/l).

Year	2004	2010
Mean	2.74	1.37
Min	0.0	0.0
Max	∞	∞
Percentile 25	0.35	0.36
Percentile 50	0.94	0.75
Percentile 75	2.53	1.57
Percentile 90	6.13	3.06
Percentile 95	10.42	4.55
Percentile 99	28.18	9.60
Percentile 99.9	85.95	22.15

4.7 Method of simulation

The preceding sections discussed the results of fitting distributions to the data sets under consideration. A common problem was large errors in the tail of the distribution, which is the area that is likely to be most important when estimating toxicological risks or the risk of exposure exceeding statutory limits. This part of the data was unaffected by censoring at the limit of detection. Conversely, some of the sets were severely censored: up to 50% in the case of lead. Any method of substitution would, therefore, lead to biases in the estimate of statistics for the majority of the population, such as the mean or median. Finally, it should be noted that the data for some of the determinands that have not been studied in detail, such as nitrate, were multi-modal, so could not be fitted by parametric distributions.

All of the data sets were large, containing 10,000–50,000 values after excluding the points below the LoD, so random sampling from the data offered a good alternative to using a distribution function, except for the problem of values below the LoD. The method chosen for the exposure simulation was the hybrid approach proposed by EFSA (2010): points were sampled from the data, but values below the LoD were substituted using the best available distribution. When the sampled value was below the LoD, the distribution function was used to convert the value to the corresponding probability of obtaining a value less than or equal to the LoD from the distribution. A random number between zero and that probability was generated and converted back to a concentration using the inverse of the distribution function (*i.e.* the quantile function). Note that repeated samples from the same point, if they occurred, would return different substituted values representing the uncertainty in the true value. Some comparisons of the exposure simulation were also performed using the approach favoured by the FSA of simulating exposure twice: once using substitution by 0 and the other using substitution by the LoD to give upper and lower bounds. The tables of summary statistics for concentration given above showed that that the distributions fitted to the data usually gave mean, median and lower percentiles in the range produced from the data using this approach.

This method could not be used for lead prior to 2004, because the data were only available as tables of class frequencies. For these years the fitted lognormal distributions were used to generate all the samples.

5. Reference data on exposure to contaminants

A comprehensive search strategy was developed to identify literature relating to dietary intakes and guidelines for the chemicals being considered within the project. A detailed set of search terms were developed as the basis for identifying relevant data from authoritative sources such as the UK Department of Health, and from peer-reviewed literature, through searches of SCOPUS (includes Medline & Embase) and CSA Illumina (Aqualine, Biological Sciences, Environment Abstracts, Environment Science and Pollution Management, Medline, Risk Abstracts, Toxline, Water Resources Abstracts).

The output of these searches was subjected to detailed consideration for relevance to the project by a risk assessor within Cranfield University's Institute of Environment and Health.

Most of the data on reference values were drawn from three authoritative UK sources (Burmaster, 1998; COT, 2003; Defra, 2010), the World Health Organisation (WHO, 1987), The Joint FAO/WHO Expert Committee on Food Additives (JECFA) (FAO, 2012) and the recommendations of the UK Food Standards Agency's Expert Group on Vitamins and Minerals (EVM). Additional information for lead came from the European Food Safety Agency (EFSA, 2010). Data on intakes from sources other than drinking water were taken from taken from the National Diet and Nutrition Survey (Bates *et al.*, 2011) and COT (2003). These are summarised in Table 32.

The reference values are expressed in several different forms:

ADI (Acceptable Daily Intake): an estimate of the amount of a substance in food or drinking water, expressed on a body-weight basis, that can be ingested daily over a lifetime without appreciable risk (standard human (WHO) = 60 kg). The ADI is listed in units of mg per kg of body weight.

Benchmark dose (BMD): a dose or concentration that produces a predetermined change in response rate of an adverse effect (called the benchmark response or BMR) compared to background.

Benchmark dose limit (BMDL): a statistical lower confidence limit on the dose or concentration at the BMD. For $BMDL_{01}$ this is at the 1% response level.

PMTDI (Provisional Maximum Tolerable Daily Intake): the endpoint used for contaminants with no cumulative properties. Its value represents permissible human exposure as a result of the natural occurrence of the substance in food and in drinking-water. In the case of trace elements that are both essential nutrients and unavoidable constituents of food, a range is expressed, the lower value representing the level of essentiality and the upper value the PMTDI.

PTWI (Provisional Tolerable Weekly Intake): an endpoint used for food contaminants such as heavy metals with cumulative properties. Its value represents permissible human weekly exposure to those contaminants unavoidably associated with the consumption of otherwise wholesome and nutritious foods.

RNI (Reference Nutrient Intake): for a vitamin or mineral, the amount of the nutrient that is sufficient for about 97% of people in the group. If the average intake of the group is at the RNI, then the risk of deficiency in the group is judged to be very small. However, if the average intake is lower than the RNI then it is possible that some of the group will have an intake below their requirement.

It is important to note that the RNI differs from the other reference values, in that it represents a level of sufficiency for an essential nutrient. The others represent upper limits of permissible or tolerable exposure. There is a subtler distinction between ADI and PTWI, which are limits of regular exposure to contaminants with cumulative properties (e.g. lead) and the PMTDI for non-cumulative minerals, which may also be essential nutrients with RNIs specified (e.g. iron and copper).

Of the six substances considered in this study, only lead is not an essential nutrient. The recommended maxima for lead have recently been reviewed. The previous PTWI (WHO, 1987) was equivalent to 3.5 μg $kg^{-1}d^{-1}$. However, EFSA now states that there is no recommended tolerable intake level as there is no evidence of thresholds for a number of critical health effects. Using the BMD approach, the 2010 EFSA opinion identified a 95th percentile lower confidence limit of the benchmark dose of 1% extra risk ($BMDL_{01}$) of 0.50 μg $kg^{-1}d^{-1}$ for developmental neurotoxicity in young children (EFSA, 2012). The report also lists cardiovascular effects and nephrotoxicity in adults as potential critical adverse health effects of lead with $BMDL_{01}$ for cardiovascular effects of 1.5 μg $kg^{-1}d^{-1}$ and $BMDL_{10}$ for nephrotoxicity of 0.635 μg $kg^{-1}d^{-1}$.

RNIs are given for iron, copper, sodium and selenium. For manganese, an FSA working group noted that (in the USA) "Insufficient data were available to set a recommended dietary amount … Adequate intakes were determined as there was insufficient information to set RDAs" (FSA, 2002). The report also stated that, "The EU Scientific Committee for Food (SCF) recommended a safe and adequate dose of 1–10 mg/Mn/person/day (SCF, 1993)." The use of the word "adequate" would imply that this is broadly similar to an RNI, but less precisely defined.

PMTDIs are given for copper and iron. There is also an EVM Safe Upper Level for total dietary intake of copper, which is about one-third of the PMTDI (COT, 2003). Both WHO and EVM give upper safe limits for selenium of similar magnitudes. The Scientific Advisory Committee on Nutrition reviewed the evidence on the health effects of salt and recommended that the average intake should be reduced to 6 g/d for adults, but stated that this was an "achievable population goal" not an optimum level (SACN, 2003). This is equivalent to a sodium intake of 2.4 g/d. FSA (2002) reported that manganese has low acute toxicity. Although there was some evidence of neurotoxic and other effects of chronic exposure, the data were inadequate to set maximum safe levels.

Table 32. Reference intake values and estimated intakes from sources other than drinking water for selected contaminants.

Contaminant	Water Supply Regulations	Intake guideline	Estimated intake from non-drinking water sources
Lead	Present 25 µg/l From 25/12/2013 10 µg/l	Young children $BMDL_{01}$ (neurotoxicity): 0.5 µg $kg^{-1}d^{-1}$ Adults $BMDL_{01}$ (cardio-vascular): 1.5 µg $kg^{-1}d^{-1}$ Adults $BMDL_{10}$ (nephrotoxicity): 0.635 µg $kg^{-1}d^{-1}$ (EFSA, 2012) **Previously** PTWI for lead (all age groups) of 25 µg/kg bw (WHO, 1987) equivalent to an ADI of approximately 3.5 µg $kg^{-1}d^{-1}$ JECFA ADI: 3.6 µg $kg^{-1}d^{-1}$ (FAO, 2012)	Mean intakes, µg $kg^{-1}d^{-1}$ (COT, 2003) Pre-school (1.5–4.5 years): 0.21–0.25 Young people (4–18 years): 0.13–0.15 Adults: 0.09–0.10
Iron	200 µg/l	RNI, mg/day (Defra, 2010) <1 year: 5.4 1–3 years: 6.9 4–6 years: 6.1 7–10 years: 8.7 11–14 years (F): 14.8 11–14 years (M): 11.3 15–18 years (F): 14.8 15–18 years (M): 11.3 19–50 years (F): 14.8 19–50 years (M): 8.7 51–64 years (F): 8.7 51–64 years (M): 8.7 JECFA PMTDI: 0.8 mg $kg^{-1}d^{-1}$ (FAO, 2012)	Mean total intake from food sources, mg/day (Bates *et al.*, 2011) Age 4–10 (F): 8.4 Age 4–10 (M): 9.1 Age 11–18 (F): 8.6 Age 11–18 (M): 10.8 Age 19–64 (F): 9.8 Age 19–64 (M): 12.0 Age 65+ (F): 9.5 Age 65+ (M): 11.3

Contaminant	Water Supply Regulations	Intake guideline	Estimated intake from non-drinking water sources
Copper	2 mg/l	RNI, mg/day (Buttriss, 2000) 0–3 months: 0.2 4–6 months: 0.3 7–9 months: 0.3 9–12 months: 0.3 1–3 years: 0.4 4–6 years: 0.6 7–10 years: 0.7 11–14 years: 0.8 15–18 years: 1.0 >18 years: 1.2 JECFA PMTDI: 0.5 mg $kg^{-1}d^{-1}$ (FAO, 2012) EVM Safe Upper Level = 0.16 mg $kg^{-1}d^{-1}$ for total dietary intake (COT, 2003)	Mean total intake from food sources, mg/day (Bates *et al.*, 2011) Age 4–10 (F): 0.79 Age 4–10 (M): 0.81 Age 11–18 (F): 0.86 Age 11–18 (M): 1.04 Age 19–64 (F): 1.06 Age 19–64 (M): 1.27 Age 65+ (F): 1.09 Age 65+ (M): 1.39
Sodium	200 mg/l	RNI, mg/day (Buttriss, 2000) < 1 year: 300 1-3 years: 500 4-6 years: 700 7-10 years: 1200 >10 years: 1600 Target level <2400 mg/d (SACN, 2003)	Mean total intake from food sources, mg/day (Bates *et al.*, 2011) Age 4-10 (F): 1863 Age 4-10 (M): 1989 Age 11-18 (F): 2009 Age 11-18 (M): 2563 Age 19-64 (F): 2029 Age 19-64 (M): 2732 Age 65+ (F): 1889 Age 65+ (M): 2393

Contaminant	Water Supply Regulations	Intake guideline	Estimated intake from non-drinking water sources
Selenium	10 µg/l	RNI, µg/day (WHO, 2001) 0-6 months: 6 7-12 months: 10 1-3 years: 17 4-6 years: 22 7-9 years: 21 10-18 years (F): 26 10-18 years (M): 32 19-65 (F): 26 19-65 (M): 34 65+ years (F): 25 65+ years (M): 33 Recommended daily intake of selenium is about 1 µg/kg bw for adults. The upper limit of the safe range proposed by the WHO is 400 mg/day determined for adults only based on epidemiological data. EVM Safe Upper Level = 450 mg/d for total dietary intake, equivalent to 7.5 mg $kg^{-1}d^{-1}$ for a 60 kg adult (COT, 2003)	Mean total intake from food sources µg/day (Bates *et al.*, 2011) Age 4-10 (F): 32.0 Age 4-10 (M): 34.0 Age 11-18 (F): 35.0 Age 11-18 (M): 44.0 Age 19-64 (F): 43.0 Age 19-64 (M): 54.0 Age 65+ (F): 41.0 Age 65+ (M): 51.0
Manganese	50 µg/l	Adequate daily intakes, mg/d (FSA, 2002) 0-6 months: 0.003 7-12 months: 0.6 1-3 years: 1.2 4-8 years: 1.5 9-13 years (M): 1.9 14-18 years (M): 2.2 9-18 years (F): 1.6 >18 years(M) 2.3 >18 years(F): 1.8 Pregnancy 2.0 Lactation: 2.6 No maximum found	Median total intake from food sources (mg/day) (FSA, 2002) 6-12 months (M&F): 1.1 1.5-2.5y (M&F): 1.0 2.5-3.5y (M&F): 1.1 3.5-4.5y (M):1.3 3.5-4.5y (F): 1.1 4-6y (M):1.68 4-6y (F):1.44 7-10y (M):1.92 7-10y (F):1.70 11-14y (M):2.10 11-14y (F):1.81 15-18y (M):2.4 15-18y (F):1.92 65-74y (M):3.23 65-74y (F):2.60

6. Simulation of exposure

The components needed for the simulation have all been described in the preceding sections. A bespoke program was written in R with a simple user interface using the *RGtk2* package to carry out the simulations and collect the results (Figure 16). Summary statistics and the CFD were displayed on the screen, with more detailed results logged to files.

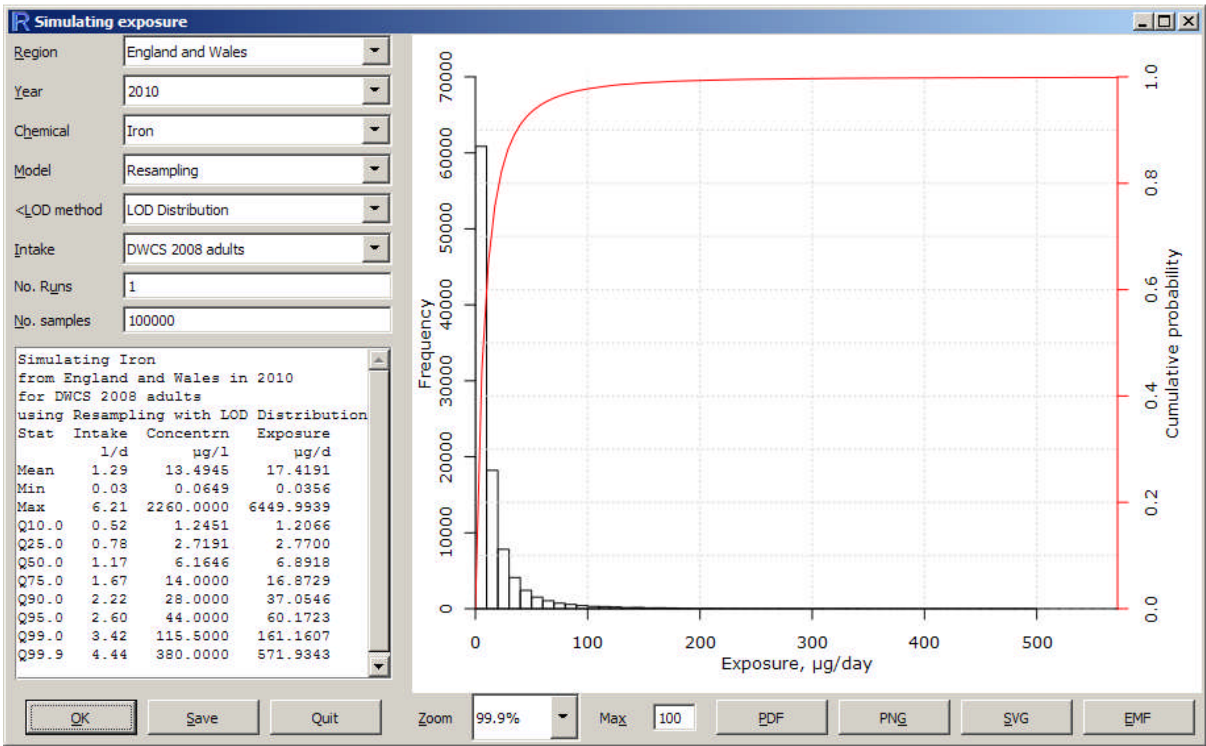


Figure 16. Graphical user interface for the exposure simulation in R.

Daily water intake was simulated from gamma distributions fitted to the different sets of survey data. The corresponding chemical concentration in the water consumed was simulated by sampling from the original concentration data, using the fitted distributions to substitute for values below the LoD. A simulation with n iterations then consisted of sampling n pairs of values, one each from the water intake and chemical concentration, and multiplying these to give the exposure. If intake is expressed in l/d and concentration is in mg/l, then the exposure is calculated in mg/cap/d. Summary statistics, such as the mean and several percentiles were derived from the results. The remaining question was to choose an appropriate value for n to give reasonably reliable estimates of the summary statistics.

The extreme percentiles of a population or simulation are always difficult to estimate reliably. Consider the case of sampling from a large population, where the true value of the 99th percentile is P99. If the sample size is 1000, the expected number of points in the sample exceeding P99 is 10. However, this number has a Poisson distribution, which also has a variance of 10 (standard deviation 3.162), so the number of such points in repeated samples will be highly variable. If the number of points in the sample with values greater than P99 (the true 99th percentile) is smaller than 10, then the 10 points with the highest values, which define the 99th percentile in the sample, will include points with values lower than P99 and will, therefore, underestimate it. Conversely, if the sample (with size 1000) contains more than 10 points above P99, then the 99th percentile will be overestimated. As the sample size increases, the standard deviation of the Poisson distribution decreases relative to its mean, so the estimation of the 99th percentile becomes more accurate.

Unfortunately, this analysis does not allow the standard error of the estimates of the percentiles to be calculated directly for the general case. However, it does help to reinforce the importance of investigating the effect of the number of iterations used in the simulation. One method that can be used to estimate the errors in statistics derived from samples is 'bootstrapping', in which the statistic of interest is calculated for many relatively small subsamples of the original sample. However, this is difficult when considering extreme percentiles: to obtain an estimate for the 99.9th percentile, each subsample would need over 1000 points. As the results were being produced by a simulation, an alternative was to test the variability of the results using multiple simulations.

The simulation was tested for each determinand with the aim of ensuring that the estimate of the 99th percentile would be reliable and the 99.9th (the 1 in 1000 level) would be reasonably reliable. The tests were performed using one intake distribution, DWCS 2008, because all the distributions were similar. Sets of 1000 runs were performed for each to calculate the mean and coefficient of variation (CoV = sd/mean) of the 99th and 99.9th percentiles. The aim was to have a CoV less than 0.1 for the 99.9th percentile and for the set of estimates of each percentile to have an unskewed (hence unbiased) distribution. It was found that 100,000 iterations were required to achieve this for those contaminants with the most highly skewed distributions (i.e. iron and lead), though 10,000 iterations were sufficient for selenium and sodium (Table 33). In the results reported below, 100,000 iterations were used for all the simulations.

To produce the results given below, each simulation was run entirely separately from the others. For each population group, the tap water intake was simulated for each substance, rather than using one set of intakes for all the substances. Similarly, each group within a survey (e.g. children 0–3 years) was simulated separately from the full population, so there was a full set of 100,000 iterations for each combination of population group and substance.

Table 33. Effect of the number of iterations on estimation of the 99th 99.9th percentiles of exposure.

Chemical	Number of iterations	Number of runs	Percentile	Mean	CoV	Shape
Iron	10,000	1000	99	162	0.05	symmetric
Iron	10,000	1000	99.9	560	0.18	skewed
Iron	100,000	100	99	162	0.02	symmetric
Iron	100,000	100	99.9	567	0.06	symmetric
Lead	10,000	1000	99	14	0.06	symmetric
Lead	10,000	1000	99.9	61	0.33	skewed
Lead	100,000	100	99	15	0.02	symmetric
Lead	100,000	100	99.9	60	0.08	symmetric
Selenium	10,000	1000	99	3.87	0.04	symmetric
Selenium	10,000	1000	99.9	7.93	0.08	symmetric
Sodium	10,000	1000	99	120	0.03	symmetric
Sodium	10,000	1000	99.9	252	0.10	symmetric

7. Simulation results

7.1 Interpreting the results

The model assumes that tap water intake and chemical concentration are independent, so each simulated point is a daily exposure event that is independent of all the others. This is most easily imagined as a set of 100,000 different individuals on a single day assuming that there is no correlation between the concentrations of chemicals delivered to individuals. Using this interpretation, the 99th percentile (for example) is an estimate of the exposure that would be exceeded by 1% of the population on a single day.

As has already been noted, it is possible that there might be some consumers or households with persistently high or low consumption, but the data did not permit investigating this. Similarly, there might be an effect of location, either through some locations or households having persistently atypical concentrations of some substances, or short-term changes affecting multiple households. There were some regional variations in the median concentrations, particularly for sodium and copper, but investigating their effects was beyond the scope of the study. The data did not permit detailed investigation of the effect of location on the higher values, but superficially there seemed to be little effect. For example, there were usually eight samples on different dates from each water supply zone (not necessarily a single household), and those that had single samples with the highest lead concentrations did not produce more than one extreme value. However, there may be effects that were not evident in the data. One instance where high concentrations may be linked to specific households is in relation to plumbing metals. The sampling method used for these metals will include a contribution from the domestic plumbing, so any such contribution will be included in the data set.

It is extremely unlikely that any individual would consistently exceed the 99th percentile, for instance, as this would require persistently high concentration and high intake. However, it is conceivable that an individual might exceed the median (or other percentile) more frequently than expected, if either the concentration or the intake was temporally correlated.

Different approaches to limits given as intakes relative to weight were used for adults and children. For adults the ADI or similar measure, where available, was converted to a daily intake using the weight of the lightest individual in the relevant group using the data from NDNS 2001. For children, when using the DW16 2011 data, because the variation in weight was greater, all the predicted intakes were converted to two values for intake relative to weight by dividing by the mean and minimum weights for the group. The use of the lightest individual represents a worst case for both adults and children, but it is particularly extreme for the youngest group. For the group of children under 4 years, the lightest was a baby under 1 year old, weighing 5 kg. This is the median weight for boys at 6 weeks or girls and 8 weeks, and the 0.4th percentile for boys at 16 weeks or girls at 22 weeks (DH, 2009), so it represents a very small proportion of the 0–3 years age group. This child's tap water intake was 0.673 l/d, which was higher than the mean for the 0–3 years age group. However, the maximum intake by babies under 1 year old was less than half that for the whole group, so it is unlikely that a baby would experience the maximum intake.

In reporting the results below, each table shows a fixed precision of 2 or 3 decimal places, determined by the smaller values. Larger values are rounded to 2–3 significant figures within the text.

7.2 Iron

7.2.1 Adults

The results of simulating exposure to iron are shown in Table 34, and a typical distribution up to the 99.9th percentile (1 in 1000) using the most recent data is shown in Figure 17. The distribution of calculated exposure was highly skewed with half the population receiving less than 7 μg/d (the median exposure), a mean exposure of 18 μg/d and 1% receiving more than 166 μg/d (the 99th percentile exposure). By increasing the resolution of the histogram (reducing the class widths), the mode was found to be non-zero, at about 1 μg/d.

Using the concentration data for 2010, the exposure results reflected the differences in water consumption between the surveys, so DWCS 2008 had the highest values for the mean, median and higher percentiles. The effect of reducing iron concentration was shown by the comparison between simulations using the 2004 and 2010 concentration results with the DWCS 2008 intake data. The mean exposure decreased from 27 μg/d to 18 μg/d, the 99th percentile (1 in 100) from 207 μg/d to 166 μg/d and the 99.9th (1 in 1000) from 704 μg/d to 533 μg/d. This is less than one-tenth of the RNI for adults (8.7 mg/d; Defra, 2010). The mean exposure via tap water was about 2% of the mean intake of 9.8–12 mg/d from other sources (Bates *et al.*, 2011). Thus, tap water does not generally make an important contribution to the required dietary intake of iron. The PMTDI is 80 mg $kg^{-1}d^{-1}$ (FAO, 2012) which is equivalent to 3120 mg/d for the lightest individual, exceeding the predicted 99.9th percentile by a factor of over 6000.

Table 34. Simulation results for exposure to iron in tap water using 2010 concentration data combined with all intake surveys containing adult respondents and 2004 concentration data combined with DWCS 2008 intake data.

Intake	DWCS 1978 all	DWCS 1995 all	DWCS 2008 adults	NDNS 2001 adults	DWCS 2008 adults
Chemical	Iron	Iron	Iron	Iron	Iron
Units	µg/d	µg/d	µg/d	µg/d	µg/d
Year	2010	2010	2010	2010	2004
Mean	12.77	15.41	17.75	15.13	26.52
Median	5.26	6.04	7.00	5.98	12.04
Upper quartile	12.55	14.75	17.10	14.54	27.14
Percentile 90	27.54	33.08	37.76	32.37	57.09
Percentile 95	44.72	54.30	61.86	53.22	89.63
Percentile 99	116.20	146.28	165.87	140.14	207.08
Percentile 99.9	399.98	496.85	532.86	462.86	704.28

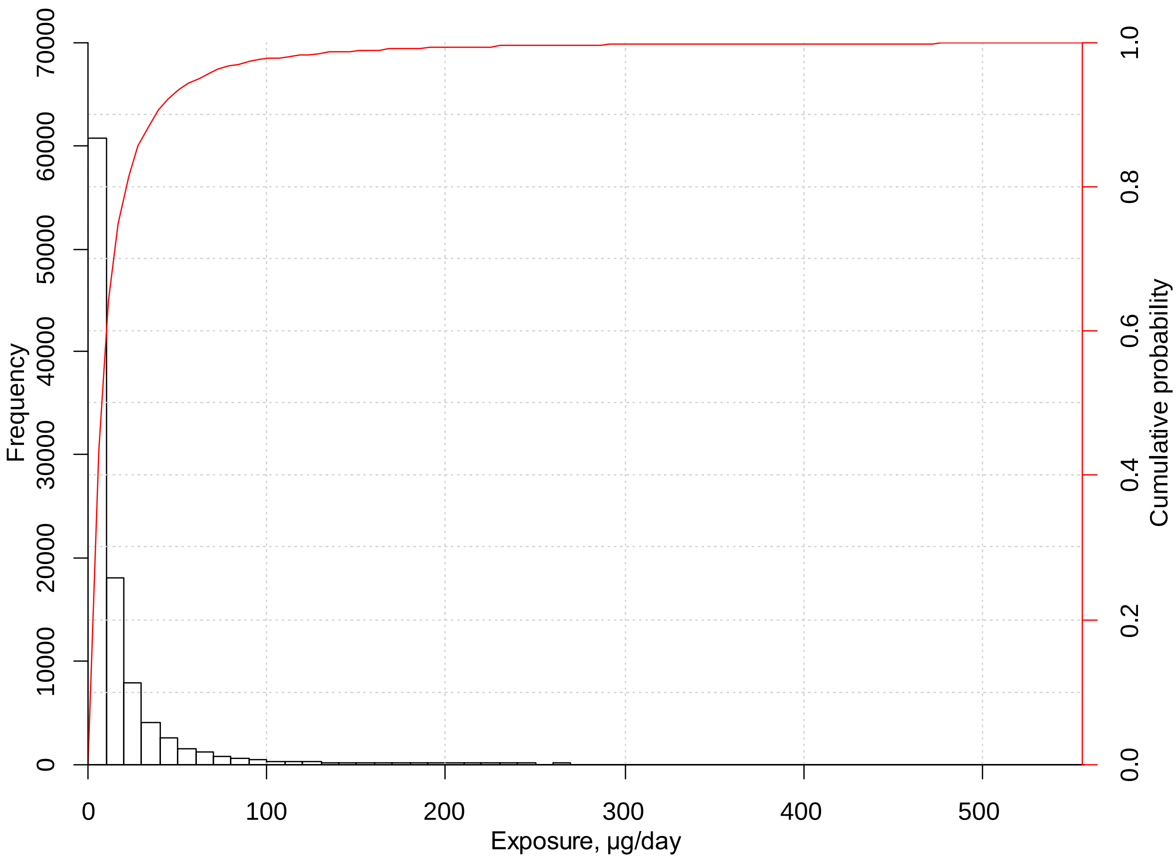


Figure 17. Simulation results (histogram and CDF) for exposure to iron in tap water using 2010 concentration data combined with DWCS 2008 intake data.

The NDNS 2001 data were used to perform separate simulations by age and sex (Table 35). The differences were due to the observed differences in tap water intake:

increasing noticeably with age and slightly higher in males. The mean, median and upper percentiles for all groups were lower than those found for all adults using DWSC 2008.

Table 35. Simulation results for exposure to iron in tap water using 2010 concentration data and NDNS survey for adults by age and sex.

Intake	NDNS 2001 19–25	NDNS 2001 25–39	NDNS 2001 40–54	NDNS 2001 54–64	NDNS 2001 females	NDNS 2001 males
Chemical	Iron	Iron	Iron	Iron	Iron	Iron
Units	μg/d	μg/d	μg/d	μg/d	μg/d	μg/d
Year	2010	2010	2010	2010	2010	2010
Mean	10.23	14.17	15.96	16.12	14.60	15.26
Median	3.99	5.58	6.45	6.56	5.73	6.16
Upper quartile	9.86	13.64	15.57	15.53	13.93	14.97
Percentile 90	21.93	30.38	34.00	33.79	31.08	33.12
Percentile 95	36.48	49.73	55.72	54.47	51.05	53.81
Percentile 99	95.88	129.39	151.64	148.04	136.72	136.73
Percentile 99.9	368.93	450.46	476.12	564.17	487.83	483.88

The effects of using different methods of substitution for values less than the LoD was explored by conducting batches of 10 runs of 100,000 iterations each using four substitutions: 0, LoD, LoD/2 and the distribution used to generate the results above (Table 36, Table 37). For each method, the range of values for all the summary statistics other than the 99.9th percentile was small, as previously found when selecting the number of iterations to use. The method of substitution affected all the statistics up to the 75th percentile; the higher percentiles were consistent across all of the methods. Substituting values below the LoD with 0 and the LoD inevitably gave values respectively below and above those for the other two methods. The range of estimates of the mean was 15.5–19.58, which is very small compared with the reference values, so any of the methods would be adequate to estimate the mean in practice. Using LoD/2 gave very similar results to using the lognormal distribution.

Table 36. Comparison of simulated exposure to iron in tap water using <LoD substitution by 0 and LoD. Mean, minimum and maximum of 10 runs with 2010 concentration data combined with DWCS 2008 intakes.

	Substitute <LoD = 0			Substitute <LoD = LoD		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
Mean	15.76	15.50	16.19	19.33	19.19	19.58
Min	0.00	0.00	0.00	0.07	0.05	0.09
Max	4968.71	3484.50	8899.66	5092.62	2704.52	6493.14
Percentile 10	0.00	0.00	0.00	2.06	2.04	2.09
Percentile 25	0.00	0.00	0.00	4.73	4.68	4.76
Percentile 50	4.30	4.20	4.40	10.18	10.13	10.23
Percentile 75	16.10	15.89	16.29	19.62	19.50	19.81
Percentile 90	37.26	36.84	37.82	37.59	37.05	38.23
Percentile 95	61.05	60.28	61.76	60.10	59.02	61.14
Percentile 99	163.21	154.99	169.29	160.32	155.69	166.83
Percentile 99.9	548.33	471.32	639.05	549.71	496.93	582.43

Table 37. Comparison of simulated exposure to iron in tap water using <LoD substitution by LoD/2 and lognormal distribution. Mean, minimum and maximum of 10 runs with 2010 concentration data combined with DWCS 2008 intakes.

	Substitute <LoD = LoD/2			Substitute <LoD = distribution		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
Mean	17.73	17.47	17.95	17.47	17.19	17.72
Min	0.03	0.02	0.05	0.02	0.02	0.03
Max	4957.95	3277.10	7737.17	4370.06	3481.96	7594.29
Percentile 10	1.44	1.42	1.46	1.20	1.18	1.22
Percentile 25	3.37	3.31	3.40	2.77	2.74	2.79
Percentile 50	7.38	7.32	7.43	6.91	6.86	6.98
Percentile 75	16.63	16.42	16.75	16.90	16.76	17.09
Percentile 90	37.28	36.93	37.47	37.41	37.06	37.90
Percentile 95	61.14	60.45	61.89	61.23	60.61	62.39
Percentile 99	162.85	159.82	166.69	163.60	161.63	167.22
Percentile 99.9	564.86	514.23	616.94	551.84	518.01	599.57

7.2.2 Children

The results of the simulation of exposure to iron for the whole sample and the sex and age groups are shown in Table 38, which includes the RNI for each age group (Defra, 2010). Because Defra (2010) distinguishes babies under 1 year old from the youngest group and splits the oldest into males and females, the ranges of RNIs are shown for these in the table. In each case, the 99.9th percentile of the simulated exposure was less than 1/20 of the RNI. The mean total intake from food sources is similar to the RNI (Bates *et al.*, 2011), so, as for adults, the simulation predicted that tap water was generally an insignificant source of intake of iron.

The results for the four weight groups are shown in Table 39 and relative to the weights of the mean and minimum weight individuals in Table 40. The PMTDI (FAO, 2012) is 800 μg $kg^{-1}d^{-1}$; the predicted relative intake for the 99.9th percentile of the <17 kg group was only 46 μg $kg^{-1}d^{-1}$ using the weight of the lightest individual.

Table 38. Simulation results for exposure to iron via tap water (μg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by sex and age with RNIs (Defra, 2010).

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
RNI				5,400–6,900	6,100	8,700	11300–14,800
Mean	7.473	6.943	7.942	6.018	6.190	7.920	9.097
Median	2.659	2.506	2.895	2.091	2.310	2.935	3.312
Upper quartile	6.908	6.456	7.400	5.529	5.822	7.448	8.445
Percentile 90	15.962	14.942	17.083	12.995	13.131	16.857	19.321
Percentile 95	26.436	24.733	28.403	21.756	21.875	27.940	32.297
Percentile 99	73.611	68.086	76.471	59.253	58.183	77.367	87.787
Percentile 99.9	280.613	245.815	281.905	209.827	215.869	275.945	336.261

Table 39. Simulation results for exposure to iron via tap water (μg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

Intake	DW16 2011 <17 kg	DW16 2011 17-26 kg	DW16 2011 26-41 kg	DW16 2011 ≥41 kg
Mean	5.867	6.677	7.608	9.580
Median	1.992	2.569	2.737	3.552
Upper quartile	5.322	6.416	7.143	8.934
Percentile 90	12.461	14.305	16.458	20.458
Percentile 95	20.978	23.418	27.758	33.923
Percentile 99	59.569	63.276	76.300	91.859
Percentile 99.9	227.943	207.496	243.615	339.217

Table 40. Simulation results for exposure to iron via tap water on a body weight basis (µg $kg^{-1}d^{-1}$) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

	Mean weight individual				Lightest individual			
Weight group, kg	< 17	17-26	26-41	≥ 41	< 17	17-26	26-41	≥ 41
Mean	0.45	0.31	0.23	0.18	1.18	0.39	0.29	0.23
Median	0.15	0.12	0.08	0.07	0.40	0.15	0.11	0.09
Upper quartile	0.41	0.30	0.22	0.17	1.07	0.38	0.27	0.22
Percentile 90	0.96	0.67	0.50	0.38	2.50	0.84	0.63	0.50
Percentile 95	1.62	1.09	0.84	0.63	4.20	1.38	1.07	0.82
Percentile 99	4.60	2.95	2.31	1.72	11.94	3.72	2.93	2.23
Percentile 99.9	17.60	9.67	7.39	6.34	45.68	12.21	9.37	8.22

7.3 Lead

7.3.1 Adults

The results of simulating exposure to lead are shown in Table 41, and a typical distribution up to the 99.9th percentile (1 in 1000) using the most recent data is shown in Figure 18. Again, the distribution was highly skewed: the mode was about 0.05 µg/d, the median was 0.25 µg/d, the mean was 1.46 µg/d and the 99th percentile was 15 µg/d.

Table 41. Simulation results for exposure to lead in tap water using 2010 concentration data with all intake surveys containing adult respondents, 2004 concentrations combined with DWCS 2008 intakes and 1994 concentrations combined with DWCS 1995 intakes.

Intake	DWCS 1978 all	DWCS 1995 all	DWCS 2008 adults	NDNS 2001 adults	DWCS 2008 adults	DWCS 1995 all
Chemical	Lead	Lead	Lead	Lead	Lead	Lead
Units	µg/d	µg/d	µg/d	µg/d	µg/d	µg/d
Year	2010	2010	2010	2010	2004	1994
Mean	1.04	1.19	1.46	1.16	2.47	4.22
Median	0.19	0.22	0.25	0.21	0.52	2.19
Upper quartile	0.56	0.66	0.75	0.63	1.78	4.81
Percentile 90	1.78	2.14	2.46	2.05	5.24	9.58
Percentile 95	3.46	4.20	4.83	3.99	8.99	14.51
Percentile 99	10.77	13.49	14.66	12.06	24.83	31.44
Percentile 99.9	45.20	52.73	53.50	49.91	127.25	*

* Not shown because the absence of detailed concentration data prevents reliable estimation

Using the concentration data for 2010, the predicted exposures reflected the differences in water consumption between the surveys, so the DWCS 2008 data yielded the highest values for the mean, median and higher percentile exposures. The effect of the decrease in lead concentrations over time was shown by the comparison between simulations using the 2004 and 2010 concentration data with the DWCS 2008 intake data. The mean exposure decreased from 2.47 µg/d to 1.46 µg/d, the 99th percentile (1 in 100) from 5.24 µg/d to 2.46 µg/d and the 99.9th (1 in 1000) from 127 µg/d to 54 µg/d. Compared with 1994, the mean exposure decreased by a factor of 4, the median by a factor of 10 and the 99th percentile by a factor of 3. The 99.9th percentile is not shown for 1994 because the absence of 'raw' data for concentration meant that the simulation used the fitted lognormal distribution, which was likely to underestimate the frequency of extreme values.

Using the lowest body weight found in NDNS 2001 (39 kg), the former ADI for lead (3.6 µg $kg^{-1}d^{-1}$; WHO, 1987) is equivalent to an intake of 137 µg/d, which is close to the 99.9th percentile for 2004, but is 2.5 times the 2010 value and about 100 times the mean exposure via tap water. The ADI is the amount of a substance that can be ingested daily over a lifetime without appreciable risk, whereas high concentrations of lead are sporadic events. Thus the mean exposure is more appropriate for comparison with the ADI than the upper percentiles. Using the lower of the BMDL values given for adults ($BMDL_{10}$ for nephrotoxicity: 0.635 µg $kg^{-1}d^{-1}$, EFSA, 2012) is equivalent to 25 kg/d for the lightest adult in the survey. This is about 20 times the predicted mean exposure and lies between the 99th and 99.9th percentiles. Again, the BMDL relates to lifetime exposure, so the risk is very low indeed.

The mean lead intake from other sources (COT, 2003) is equivalent to 3.9 µg/d for the lightest adult, which is more than double the mean exposure via tap water. The intake from other sources for a 'standard' 60 kg adult is equivalent to 6.0 µg/d, which is greater than the 95th percentile of the exposure via tap water.

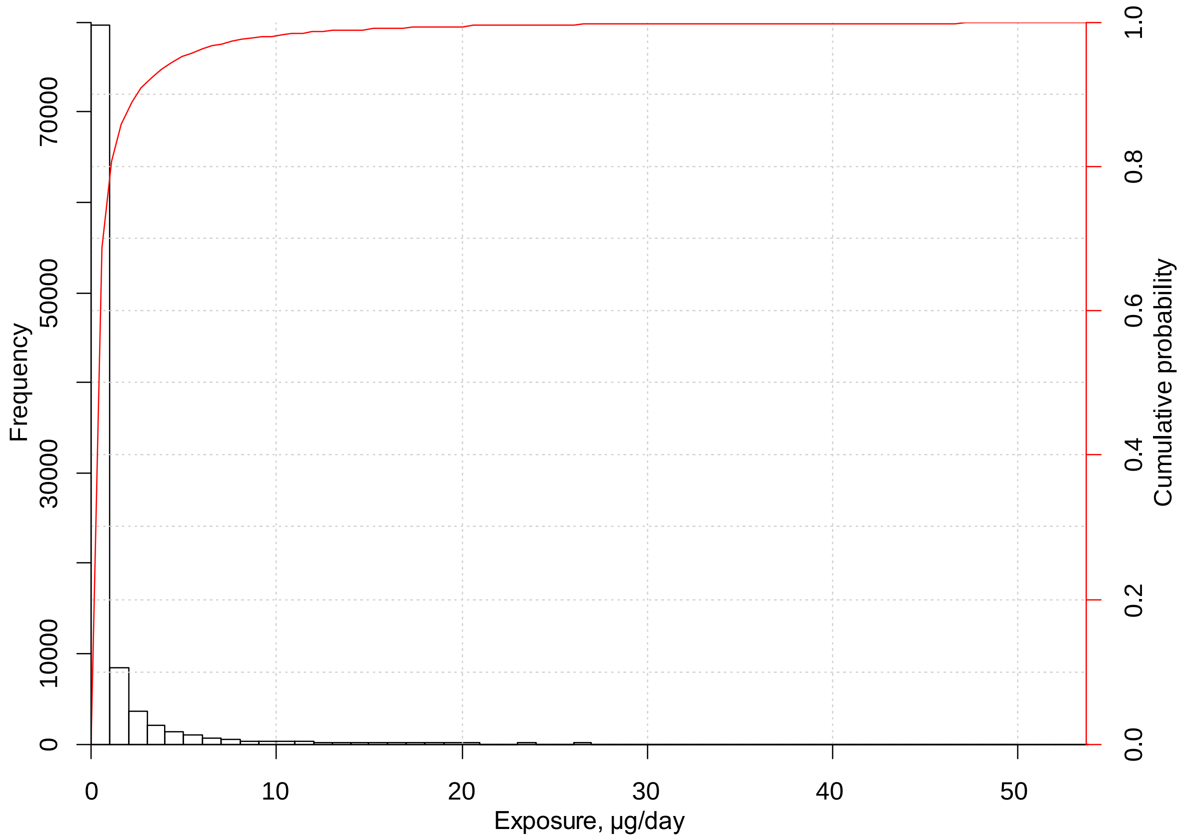


Figure 18. Simulation results (histogram and CDF) for exposure to lead in tap water using 2010 concentration data combined with DWCS 2008 intake data.

Using the NDNS 2001 data analysed by age and sex, the predicted exposure (Table 42) shows the same pattern as for iron. Due to the generally lower tap water intakes found in that survey, the mean, median and upper percentiles of exposure for all groups are lower than those found using DWCS 2008 intake data. The table includes an estimate of exposure which is equivalent to the ADI multiplied by the body mass of the lightest individual. In each case this ADI-equivalent exposure (ADIex) exceeded the 99.9th percentile simulated exposure by a factor greater than 2.

The effect of the different methods of substitution for values less than the LoD was tested and gave a similar pattern of results to those for iron (Table 43, Table 44). All the percentiles up to the 50th (the median) are strongly influenced by the substitution used. Substituting 0 for the LoD, these percentiles are all very close to 0, the results obtained using half the LoD and the values from the distribution are similar, and these are about half those resulting from substituting by the LoD. The effect on the 75th percentile (the upper quartile) is small, and there is no effect on the 90th and higher percentiles. The effect on the mean is comparatively small, because it is dominated by relatively few high values. On the basis of these results, the use of the half-LoD substitution would appear to be adequate for most purposes, especially as it is normally high intakes that are of most concern.

Table 42. Simulation results for exposure to lead in tap water using 2010 concentration data and NDNS survey for adults by age and sex. Exposure equivalent to the product of the ADI and the minimum weight for each group (ADIex) is also shown.

Intake	NDNS 2001 19–25	NDNS 2001 25–39	NDNS 2001 40–54	NDNS 2001 54–64	NDNS 2001 females	NDNS 2001 males
Chemical	Lead	Lead	Lead	Lead	Lead	Lead
Units	µg/d	µg/d	µg/d	µg/d	µg/d	µg/d
Year	2010	2010	2010	2010	2010	2010
Mean	0.80	1.10	1.32	1.44	1.19	1.24
Median	0.15	0.20	0.23	0.24	0.21	0.22
Upper quartile	0.44	0.61	0.71	0.70	0.62	0.66
Percentile 90	1.41	1.98	2.26	2.28	2.01	2.18
Percentile 95	2.76	3.84	4.34	4.38	3.97	4.21
Percentile 99	8.45	11.65	13.21	13.36	12.43	12.44
Percentile 99.9	36.35	44.80	55.57	57.05	57.63	50.71
ADIex	149.8	154.4	149.1	136.5	136.5	136.5

Table 43. Comparison of simulated exposure to lead in tap water using <LoD substitution by 0 and LoD. Mean, minimum and maximum of 10 runs with 2010 concentration data combined with DWCS 2008 intakes .

	Substitute <LoD = 0			**Substitute <LoD = LoD**		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
Mean	1.32	1.22	1.44	1.59	1.47	1.73
Min	0.00	0.00	0.00	0.00	0.00	0.00
Max	3643.95	2158.77	4854.60	3803.76	2559.68	5009.85
Percentile 10	0.00	0.00	0.00	0.10	0.10	0.10
Percentile 25	0.00	0.00	0.00	0.22	0.22	0.22
Percentile 50	0.01	0.00	0.02	0.48	0.48	0.49
Percentile 75	0.68	0.67	0.69	1.02	1.01	1.03
Percentile 90	2.43	2.38	2.48	2.49	2.46	2.53
Percentile 95	4.78	4.71	4.89	4.75	4.67	4.84
Percentile 99	14.65	14.16	14.81	14.60	14.28	14.87
Percentile 99.9	58.21	51.26	64.44	61.80	54.90	70.06

Table 44. Comparison of simulated exposure to lead in tap water using <LoD substitution by LoD/2 and lognormal distribution. Mean, minimum and maximum of 10 runs with 2010 concentration data combined with DWCS 2008 intakes.

	Substitute <LoD = LoD/2			Substitute <LoD = distribution		
	Mean	Min	Max	Mean	Min	Max
Mean	1.42	1.34	1.53	1.40	1.28	1.47
Min	0.00	0.00	0.00	0.00	0.00	0.00
Max	3567.74	2116.81	5834.72	3167.04	1838.29	4623.04
Percentile 10	0.07	0.06	0.07	0.04	0.04	0.04
Percentile 25	0.14	0.14	0.14	0.10	0.10	0.10
Percentile 50	0.31	0.31	0.31	0.25	0.25	0.25
Percentile 75	0.77	0.76	0.78	0.75	0.74	0.75
Percentile 90	2.43	2.41	2.47	2.43	2.38	2.45
Percentile 95	4.76	4.70	4.83	4.77	4.70	4.86
Percentile 99	14.61	14.35	14.93	14.68	14.39	15.06
Percentile 99.9	58.72	54.62	68.30	58.50	49.20	66.77

7.3.2 Children

The results of the simulation of exposure to lead for the whole sample and the sex and age groups using the concentration data for 2004 and 2010 are shown in Table 45 and Table 46. As for adults, this shows a reduction of about 40% in most of the exposure statistics and a slightly larger reduction in the 99.9th percentile.

Table 45. Simulation results for exposure to lead via tap water (µg/d) using 2004 concentration data and intake from DW16 2011 for under-16s by sex and age.

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
Mean	1.001	0.961	1.086	0.852	0.849	1.075	1.321
Median	0.203	0.187	0.217	0.158	0.173	0.223	0.251
Upper quartile	0.709	0.655	0.767	0.555	0.591	0.758	0.865
Percentile 90	2.172	2.019	2.323	1.742	1.785	2.308	2.626
Percentile 95	3.966	3.706	4.154	3.205	3.162	4.153	4.697
Percentile 99	11.107	10.438	11.367	9.387	8.771	11.960	13.397
Percentile 99.9	48.657	41.438	44.208	39.840	39.036	49.115	71.081

Table 46. Simulation results for exposure to lead via tap water (µg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by sex and age.

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
Mean	0.575	0.553	0.688	0.458	0.510	0.640	0.746
Median	0.098	0.092	0.107	0.077	0.085	0.107	0.123
Upper quartile	0.300	0.284	0.327	0.241	0.258	0.330	0.375
Percentile 90	0.994	0.930	1.088	0.789	0.844	1.073	1.216
Percentile 95	2.001	1.866	2.154	1.585	1.655	2.123	2.442
Percentile 99	6.328	5.853	6.922	5.154	5.244	6.933	7.848
Percentile 99.9	26.987	25.988	28.185	20.672	20.444	30.401	32.308

The results of the simulation for the four weight groups using the concentration data for 2010 are shown in Table 47 and relative to the weights of the mean and minimum weight individuals in Table 48. The former ADI for lead was approximately 3.5 µg $kg^{-1}d^{-1}$ (WHO, 1987). The predicted mean exposure was less than 1% of this for most groups, and less than 3% of it for the extreme case of using the intake distribution for the <17 kg group with the weight of the lightest individual (5 kg). The predicted 99.9th percentile was less than half the ADI for most groups; in the worst case it exceeded it slightly. Once again it should be emphasized that the ADI relates to long-term intake, so the mean exposure is the most appropriate comparison.

The predicted mean exposure was less than 6% of the $BMDL_{01}$ (0.5 µg $kg^{-1}d^{-1}$; EFSA, 2010) for most groups and less than 20% in the worst case. The mean exposure in the worst case was substantially less than the estimated intake from other sources, which was 0.21–0.25 µg $kg^{-1}d^{-1}$ for ages 1.5–4.5 years and 0.13–0.15 µg $kg^{-1}d^{-1}$ for ages 4–18 years (COT, 2003). These intakes from other sources are close to the predicted 95th percentile exposure via tap water. The $BMDL_{01}$ was between the predicted 99th and 99.9th percentiles for most groups and between the 95th and 99th percentiles in the worst case. Thus the risk of persistently exceeding this level is very small.

Table 47. Simulation results for exposure to lead via tap water (µg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

Intake	DW16 2011 <17 kg	DW16 2011 17-26 kg	DW16 2011 26-41 kg	DW16 2011 ≥41 kg
Mean	0.440	0.543	0.630	0.816
Median	0.074	0.093	0.100	0.130
Upper quartile	0.232	0.284	0.310	0.397
Percentile 90	0.758	0.923	1.006	1.297
Percentile 95	1.556	1.831	2.025	2.600
Percentile 99	5.086	5.666	6.464	8.213
Percentile 99.9	19.756	25.240	28.090	32.018

Table 48. Simulation results for exposure to lead via tap water on a body weight basis (μg $kg^{-1}d^{-1}$) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

	Mean weight individual				Lightest individual			
Weight group, kg	< 17	17-26	26-41	≥ 41	< 17	17-26	26-41	≥ 41
Mean	0.03	0.03	0.02	0.02	0.09	0.03	0.02	0.02
Median	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00
Upper quartile	0.02	0.01	0.01	0.01	0.05	0.02	0.01	0.01
Percentile 90	0.06	0.04	0.03	0.02	0.15	0.05	0.04	0.03
Percentile 95	0.12	0.09	0.06	0.05	0.31	0.11	0.08	0.06
Percentile 99	0.39	0.26	0.20	0.15	1.02	0.33	0.25	0.20
Percentile 99.9	1.53	1.18	0.85	0.60	3.96	1.48	1.08	0.78

7.4 Selenium

7.4.1 Adults

The results of simulating exposure to selenium are shown in Table 49, and a typical distribution up to the 99.9th percentile (1 in 1000) is shown in Figure 19. The distribution was much less skewed than those for iron, lead and copper. The mode could not be distinguished from zero even at high resolutions of the histogram. Exposures appear to have decreased by about one-third since 2004. The RNI for adults is 25–34 μg/d depending on age and sex (WHO, 2001), which is exceed by the mean intake from other sources (Bates *et al.*, 2011). The predicted mean exposure from tap water was less than one-fortieth of this, so tap water generally makes a very small contribution to the required intake of selenium. The EVM Safe Upper Level is about 10 times the RNI (COT, 2003), so the risk from tap water intake is negligible.

Table 49. Simulation results for exposure to selenium in tap water using 2010 concentration data combined with all intake surveys containing adult respondents and 2004 concentration data combined with DWCS 2008 intake data.

Intake	DWCS 1978 all	DWCS 1995 all	DWCS 2008 adults	NDNS 2001 adults	DWCS 2008 adults
Chemical	Selenium	Selenium	Selenium	Selenium	Selenium
Units	μg/d	μg/d	μg/d	μg/d	μg/d
Year	2010	2010	2010	2010	2004
Mean	0.45	0.54	0.61	0.52	1.00
Median	0.27	0.31	0.36	0.30	0.54
Upper quartile	0.57	0.68	0.78	0.66	1.22
Percentile 90	1.03	1.27	1.43	1.21	2.44
Percentile 95	1.44	1.80	2.02	1.69	3.56
Percentile 99	2.79	3.45	3.93	3.25	6.68
Percentile 99.9	5.96	7.18	8.05	6.62	12.51

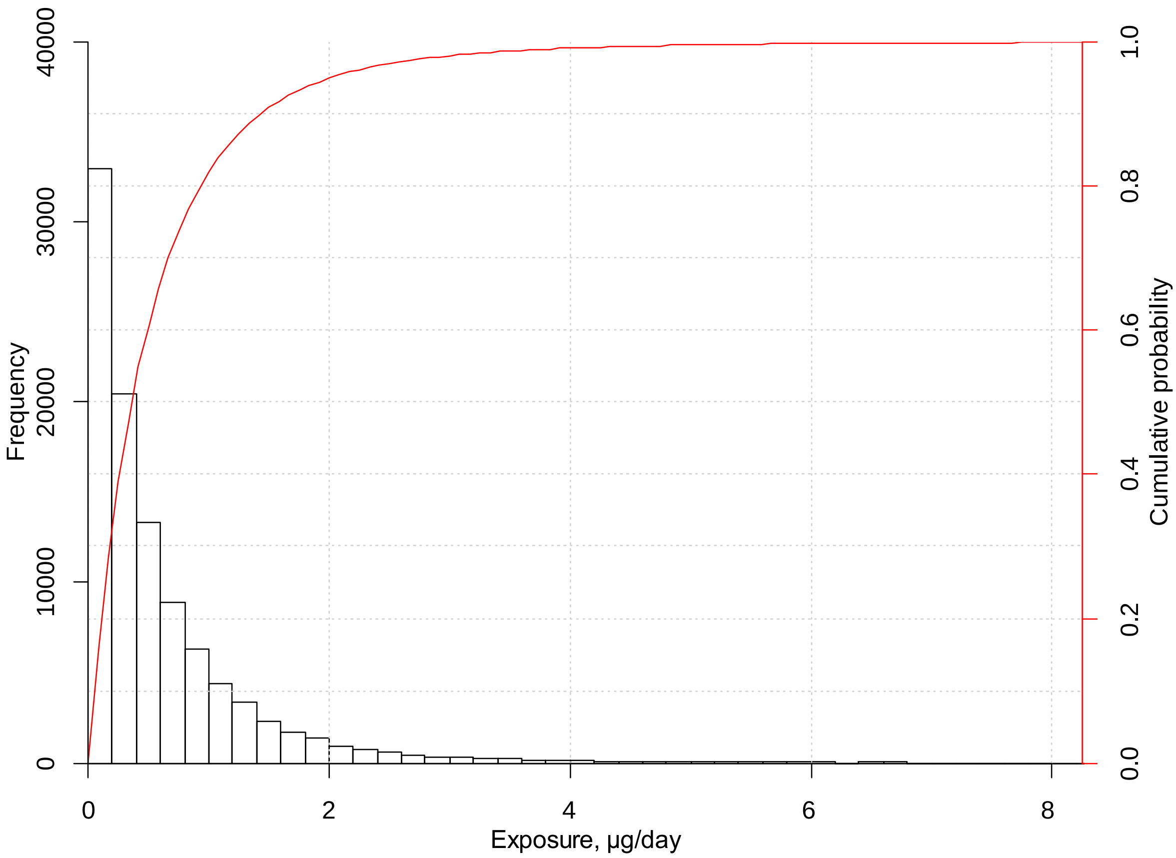


Figure 19. Simulation results (histogram and CDF) for exposure to selenium in tap water using 2010 concentration data combined with DWCS 2008 intake data.

Simulations using the NDNS 2001 intake data analysed by age and sex show the same pattern as before resulting from the variations in tap water intake between the groups, all of which were lower for this survey than for DWCS 2008 (Table 50).

Table 50. Simulation results for exposure to selenium in tap water using 2010 concentration data and NDNS survey data for adults by age and sex.

Intake	NDNS 2001 19–25	NDNS 2001 25–39	NDNS 2001 40–54	NDNS 2001 54–64	NDNS 2001 females	NDNS 2001 males
Chemical	Selenium	Selenium	Selenium	Selenium	Selenium	Selenium
Units	μg/d	μg/d	μg/d	μg/d	μg/d	μg/d
Year	2010	2010	2010	2010	2010	2010
Mean	0.36	0.50	0.56	0.56	0.51	0.54
Median	0.21	0.29	0.33	0.34	0.29	0.32
Upper quartile	0.46	0.63	0.72	0.72	0.64	0.69
Percentile 90	0.85	1.16	1.30	1.29	1.18	1.26
Percentile 95	1.21	1.63	1.83	1.79	1.67	1.77
Percentile 99	2.35	3.19	3.57	3.46	3.25	3.37
Percentile 99.9	4.85	6.87	7.60	7.08	6.97	6.84
RNI					26	33

The effect of the different methods of substitution for values less than the LoD was tested and gave similar results (data not shown) to those for iron, but with generally smaller differences between the methods, especially at the 75th percentile, due to the lower proportion of values below the LoD for selenium.

7.4.2 Children

The results of the simulation of exposure to selenium for the whole sample and the sex and age groups are shown in Table 51, which includes the RNI for each age group (WHO, 2001). Ranges of RNIs are shown for the youngest and oldest groups, as explained in Section 7.2.2. The 99.9th percentile of the predicted exposure was consistently less than one-fifth of the RNI for all but the youngest group, where it was about half of the minimum RNI. The predicted mean intake was about 1/30 of the minimum RNI for the youngest group. The mean total intake from food sources usually slightly exceeds the RNI (Bates *et al.*, 2011), so drinking water is generally predicted to make a minor contribution to the required intake, though it may be significant in a few cases.

The results for the four weight groups are shown in Table 52 and relative to the weights of the mean and minimum weight individuals in Table 53. None of the reference data sources give safe intake limits for children, but the EVM safe upper level for adults is equivalent to 7.5 mg $kg^{-1}d^{-1}$ (COT, 2003); the 99.9th percentile for the lightest group based on the weight of the lightest individual was less than 1/10,000 of this.

Table 51. Simulation results for exposure to selenium via tap water (µg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by sex and age with RNIs (WHO, 2001).

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
RNI				6–17	22	21	26–32
Mean	0.259	0.242	0.277	0.210	0.212	0.278	0.317
Median	0.136	0.127	0.146	0.105	0.117	0.151	0.170
Upper quartile	0.319	0.298	0.340	0.253	0.264	0.344	0.390
Percentile 90	0.621	0.580	0.662	0.511	0.502	0.662	0.755
Percentile 95	0.907	0.843	0.966	0.762	0.721	0.955	1.083
Percentile 99	1.780	1.671	1.908	1.541	1.430	1.878	2.190
Percentile 99.9	3.808	3.680	3.988	3.373	2.940	4.081	4.647

Table 52. Simulation results for exposure to selenium via tap water (µg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

Intake	DW16 2011 <17 kg	DW16 2011 17-26 kg	DW16 2011 26-41 kg	DW16 2011 ≥41 kg
Mean	0.202	0.237	0.268	0.333
Median	0.100	0.133	0.138	0.181
Upper quartile	0.246	0.297	0.327	0.414
Percentile 90	0.493	0.557	0.642	0.791
Percentile 95	0.722	0.798	0.943	1.137
Percentile 99	1.453	1.526	1.909	2.276
Percentile 99.9	3.216	3.255	4.132	4.803

Table 53. Simulation results for exposure to selenium via tap water on a body weight basis (µg $kg^{-1}d^{-1}$) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

	Mean weight individual				Lightest individual			
Weight group, kg	< 17	17-26	26-41	≥ 41	< 17	17-26	26-41	≥ 41
Mean	0.02	0.01	0.01	0.01	0.04	0.01	0.01	0.01
Median	0.01	0.01	0.00	0.00	0.02	0.01	0.01	0.00
Upper quartile	0.02	0.01	0.01	0.01	0.05	0.02	0.01	0.01
Percentile 90	0.04	0.03	0.02	0.01	0.10	0.03	0.02	0.02
Percentile 95	0.06	0.04	0.03	0.02	0.14	0.05	0.04	0.03
Percentile 99	0.11	0.07	0.06	0.04	0.29	0.09	0.07	0.06
Percentile 99.9	0.25	0.15	0.13	0.09	0.64	0.19	0.16	0.12

7.5 Sodium

7.5.1 Adults

The results of simulating exposure to sodium are shown in Table 54, and a typical distribution up to the 99.9th percentile (1 in 1000) is shown in Figure 20. This figure shows a much less skewed distribution than for the other determinands, with the mode clearly further from 0 as a proportion of the median: it was found to be 6–7 mg/d. Sodium was the only ion for which exposure was calculated to have increased over time, due to the increase in its concentration in tap water. However, the mean exposure via tap water using DWCS intakes was still less than 2% of the RNI for adults (1600 mg/d; Buttriss, 2000), which is exceeded by the mean intake through other sources (Bates *et al.*, 2011), so tap water generally makes a minor contribution to the intake of sodium. The 99.9th percentile was 16% of the RNI or 10% of the recommended maximum intake (SACN, 2003).

Table 54. Simulation results for exposure to sodium in tap water using 2010 concentration data combined with all intake surveys containing adult respondents and 2004 concentration data combined with DWCS 2008 intake data.

Intake	DWCS 1978 all	DWCS 1995 all	DWCS 2008 adults	NDNS 2001 adults	DWCS 2008 adults
Chemical	Sodium	Sodium	Sodium	Sodium	Sodium
Units	mg/d	mg/d	mg/d	mg/d	mg/d
Year	2010	2010	2010	2010	2004
Mean	18.76	22.46	25.42	21.56	23.77
Median	13.26	15.24	17.55	14.91	15.93
Upper quartile	24.10	28.61	32.42	27.53	30.04
Percentile 90	39.34	48.48	54.29	46.01	51.24
Percentile 95	51.58	64.24	71.87	60.83	69.16
Percentile 99	85.35	112.75	122.41	102.22	117.43
Percentile 99.9	185.63	239.01	257.14	220.68	237.83

The analysis by age and sex using the NDNS 2001 intake data showed the same pattern as for the other chemicals due to the variations in tap water intake (Table 55). The mean exposure of females was 1% of the mean intake from other sources and that for males was 0.8%. The 99.9th percentile in each case was about 10 times the mean.

Different methods of substitution for values less than the LoD were not tested, because there were very few such values for sodium.

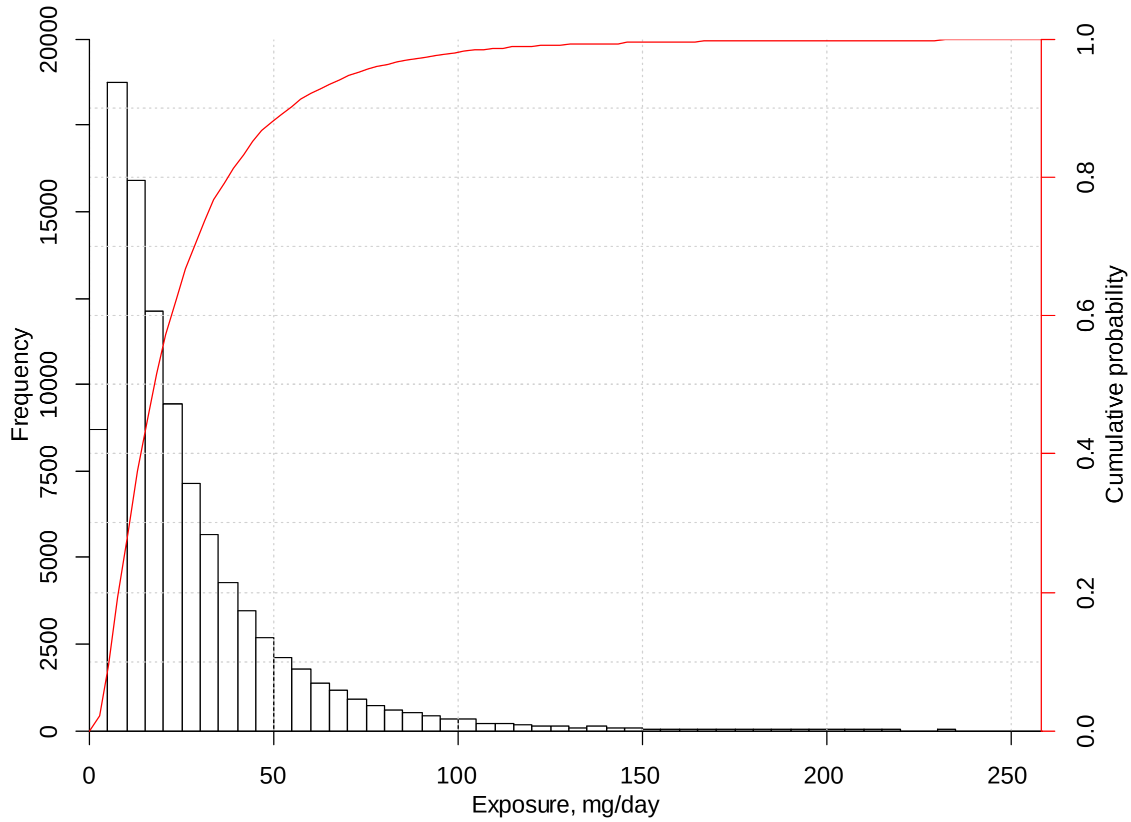


Figure 20. Simulation results (histogram and CDF) for exposure to sodium in tap water using 2010 concentration data combined with DWCS 2008 intake data.

Table 55. Simulation results for exposure to sodium in tap water using 2010 concentration data and NDNS survey for adults by age and sex.

Intake	NDNS 2001 19–25	NDNS 2001 25–39	NDNS 2001 40–54	NDNS 2001 54–64	NDNS 2001 females	NDNS 2001 males
Chemical	Sodium	Sodium	Sodium	Sodium	Sodium	Sodium
Units	mg/d	mg/d	mg/d	mg/d	mg/d	mg/d
Year	2010	2010	2010	2010	2010	2010
Mean	14.94	20.60	23.14	23.20	21.00	22.37
Median	10.16	14.18	16.36	16.54	14.51	15.68
Upper quartile	19.06	26.16	29.63	29.77	26.96	28.50
Percentile 90	32.19	43.87	48.61	48.63	44.67	47.34
Percentile 95	43.34	58.67	63.82	63.49	59.57	62.61
Percentile 99	73.82	99.88	107.33	104.55	100.29	105.04
Percentile 99.9	148.72	218.34	225.53	221.91	210.09	222.99
RNI	1600	1600	1600	1600	1600	1600
Other sources					2029	2732

7.5.2 Children

The results of the simulation of exposure to sodium for the whole sample and the sex and age groups are shown in Table 56, which includes the RNI for each age group (Buttriss, 2000). Ranges of RNIs are shown for the youngest and oldest classes, as explained in Section 7.2.2. The 99.9th percentile of the predicted exposure was about one-third of the RNI for the youngest group, decreasing to less than one-tenth for the oldest. The predicted mean intake was less than 1/30 of the RNI for the youngest group and much less for the other groups. The mean total intake from food sources consistently exceeds the RNI by a substantial margin (Bates *et al.*, 2011), so drinking water is generally predicted to be a minor source of sodium. The recommended maximum intake of sodium is about 1½ times the RNI, so it is unlikely to be exceeded.

The results for the four weight groups are shown in Table 57 and relative to the weights of the mean and minimum weight individuals in Table 58. None of the reference data sources gave values relative to body weight. As expected, the relative intake was highest for the lightest group.

Table 56. Simulation results for exposure to sodium via tap water (mg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by sex and age with RNIs (Buttriss, 2000).

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
RNI				300–500	700	1200	1600
Mean	10.753	9.974	11.496	8.694	8.846	11.484	13.159
Median	6.814	6.342	7.392	5.366	5.892	7.533	8.529
Upper quartile	13.504	12.586	14.585	10.938	11.178	14.562	16.592
Percentile 90	23.992	22.384	25.452	19.636	19.227	25.144	28.974
Percentile 95	33.111	30.645	34.988	27.398	26.081	34.213	39.898
Percentile 99	58.803	53.709	61.826	49.702	45.347	60.199	70.089
Percentile 99.9	125.094	114.192	126.317	99.813	102.529	127.330	148.819

Table 57. Simulation results for exposure to sodium via tap water (mg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

Intake	DW16 2011 <17 kg	DW16 2011 17-26 kg	DW16 2011 26-41 kg	DW16 2011 ≥41 kg
Mean	8.460	9.705	11.087	13.896
Median	5.131	6.529	7.046	9.087
Upper quartile	10.562	12.350	14.008	17.516
Percentile 90	19.268	21.063	24.898	30.775
Percentile 95	27.001	28.395	34.093	41.859
Percentile 99	49.010	48.428	60.108	72.601
Percentile 99.9	105.495	100.713	133.329	145.235

Table 58. Simulation results for exposure to sodium via tap water on a body weight basis (mg kg^{-1}d^{-1}) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

	Mean weight individual				Lightest individual			
Weight group, kg	< 17	17-26	26-41	≥ 41	< 17	17-26	26-41	≥ 41
Mean	0.65	0.45	0.34	0.26	1.70	0.57	0.43	0.34
Median	0.40	0.30	0.21	0.17	1.03	0.38	0.27	0.22
Upper quartile	0.82	0.58	0.42	0.33	2.12	0.73	0.54	0.42
Percentile 90	1.49	0.98	0.76	0.57	3.86	1.24	0.96	0.75
Percentile 95	2.09	1.32	1.03	0.78	5.41	1.67	1.31	1.01
Percentile 99	3.78	2.26	1.82	1.36	9.82	2.85	2.31	1.76
Percentile 99.9	8.15	4.69	4.04	2.71	21.14	5.92	5.13	3.52

7.6 Manganese

7.6.1 Adults

The results of simulating exposure to manganese are shown in Table 59, and a typical distribution up to the 99.9th percentile (1 in 1000) using the most recent data is shown in Figure 21. The distribution of calculated exposure was highly skewed with half the population receiving less than 0.88 µg/d (the median exposure), a mean exposure of 1.9 µg/d and 1% receiving more than 16 µg/d (the 99th percentile exposure). By increasing the resolution of the histogram (reducing the class widths), the mode was found to be non-zero, at about 0.2 µg/d.

Using the concentration data for 2010, the exposure results reflected the differences in water consumption between the surveys, so DWCS 2008 had the highest values for the mean, median and higher percentiles. The effect of reducing manganese concentration was shown by the comparison between simulations using the 2004 and 2010 concentration results with the DWCS 2008 intake data. The mean exposure decreased from 3.2 µg/d to 1.9 µg/d, the 99th percentile (1 in 100) from 28 µg/d to 16 µg/d and the 99.9th (1 in 1000) from 103 µg/d to 61 µg/d. These intakes are all very small compared with the adequate daily intakes for adults and typical intake via food, which are both around 2000 µg/d (FSA, 2002).

The NDNS 2001 data were used to perform separate simulations by age and sex (Table 60). The differences were due to the observed differences in tap water intake: increasing noticeably with age and slightly higher in males. The mean, median and upper percentiles for all groups were lower than those found for all adults using DWCS 2008.

Table 59. Simulation results for exposure to manganese in tap water using 2010 concentration data combined with all intake surveys containing adult respondents and 2004 concentration data combined with DWCS 2008 intake data.

Intake	DWCS 1978 all	DWCS 1995 all	DWCS 2008 adults	NDNS 2001 adults	DWCS 2008 adults
Chemical	Manganese	Manganese	Manganese	Manganese	Manganese
Units	µg/d	µg/d	µg/d	µg/d	µg/d
Year	2010	2010	2010	2010	2004
Mean	1.403	1.714	1.913	1.655	3.203
Median	0.660	0.762	0.876	0.749	1.250
Upper quartile	1.386	1.664	1.882	1.614	3.071
Percentile 90	2.827	3.454	3.900	3.341	7.254
Percentile 95	4.555	5.598	6.290	5.364	11.742
Percentile 99	11.801	15.290	16.075	14.180	27.854
Percentile 99.9	43.706	54.967	61.274	55.728	103.084

Table 60. Simulation results for exposure to manganese in tap water using 2010 concentration data and NDNS survey for adults by age and sex.

Intake	NDNS 2001 19–25	NDNS 2001 25–39	NDNS 2001 40–54	NDNS 2001 54–64	NDNS 2001 females	NDNS 2001 males
Chemical	Mn	Mn	Mn	Mn	Mn	Mn
Units	µg/d	µg/d	µg/d	µg/d	µg/d	µg/d
Year	2010	2010	2010	2010	2010	2010
Mean	1.129	1.536	1.768	1.810	1.601	1.730
Median	0.506	0.704	0.815	0.830	0.720	0.779
Upper quartile	1.105	1.515	1.735	1.740	1.552	1.669
Percentile 90	2.297	3.154	3.555	3.569	3.253	3.451
Percentile 95	3.718	5.071	5.767	5.752	5.216	5.551
Percentile 99	10.042	13.252	14.963	14.824	13.780	14.951
Percentile 99.9	34.522	44.642	54.345	60.698	51.289	56.883

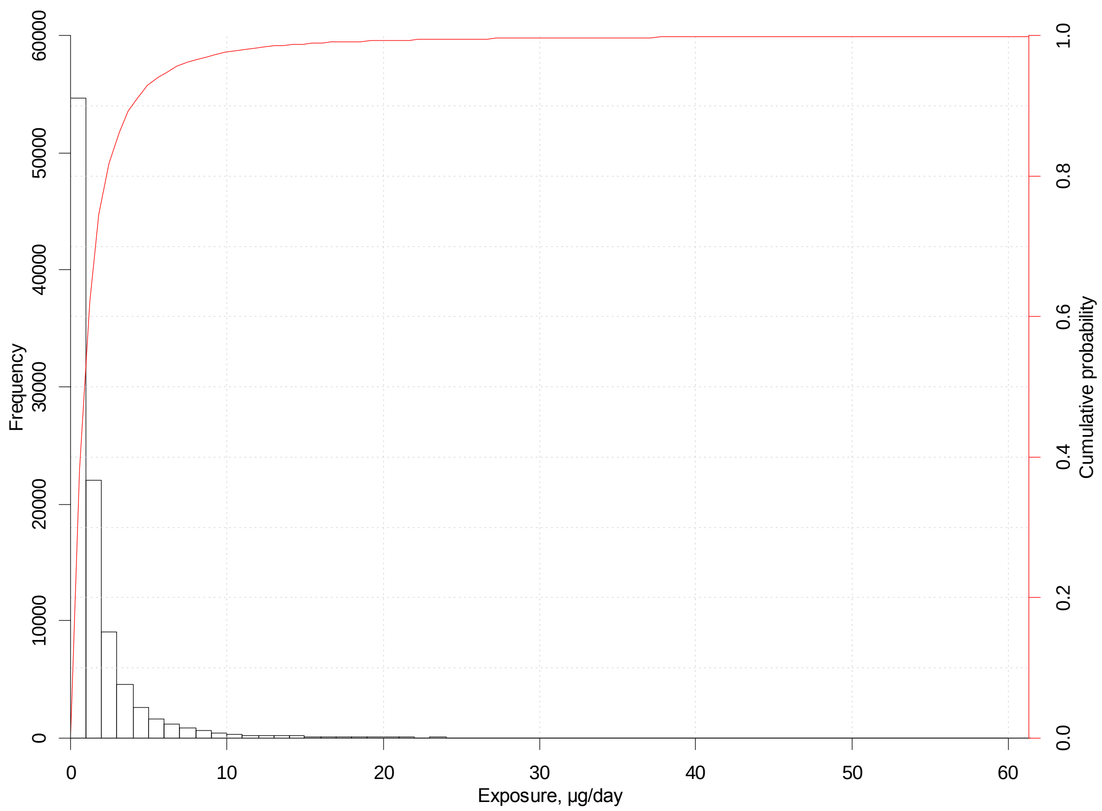


Figure 21. Simulation results (histogram and CDF) for exposure to manganese in tap water using 2010 concentration data combined with DWCS 2008 intake data.

7.6.2 Children

The results of the simulation of exposure to manganese for children under 16 using the DW16 2011 intake data are shown in Table 61 for the whole sample and the sex and age groups. It includes the adequate daily intake for each age group (FSA, 2002). Because FSA (2002) uses different age groups and splits the older groups into males and females, the ranges of adequate daily intakes are shown for these in the table. The predicted 99th percentile for the 0–3 years group exceeded the adequate daily intake of 3 µg/d for babies under 6 months, but not the much higher adequate daily intake of 600 µg/d (thirty times the predicted 99.9th percentile) for 7–12 months. Based on the recorded weights, the lightest individuals in this group were probably under 6 months old. However, as noted in Section 2.10, the maximum intake for babies under 1 year was 0.86 l/d, compared with 1.99 l/d for all 0–3 years, so the maximum exposure is likely to be proportionally lower for babies. It is therefore possible that some babies may receive or exceed the adequate daily intake on some occasions, but the predicted mean intake is still much lower than this. The adequate daily intake relates to the requirement for manganese as a nutrient. Manganese has low acute toxicity (FSA, 2002), and the highest predicted exposures were far less than the "safe and adequate" range of 1–10 mg/d proposed by SCF (1993), though this does not relate specifically to babies. For all the other age groups, the simulated exposure was insignificant compared with the adequate daily intake or the intake from food, which is of similar magnitude to the adequate daily intake.

The results for the weight groups are given in Table 62 and relative to weight in Table 63, but FSA (2002) gives no upper limits with which to compare them. Given the results above, it is highly unlikely that the intake of most individuals comes close either to requirements or to exceeding safe levels.

Table 61. Simulation results for exposure to manganese via tap water using 2010 concentration data and intake from DW16 2011 for children under 16 by sex and age with adequate daily intakes (FSA, 2002).

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
Units	µg/d	µg/d	µg/d	µg/d	µg/d	µg/d	µg/d
Adequate DI				3–1200	1500	1500–1900	1600–2200
Mean	0.807	0.753	0.878	0.660	0.670	0.859	0.994
Median	0.337	0.314	0.366	0.267	0.290	0.373	0.422
Upper quartile	0.780	0.724	0.834	0.627	0.646	0.842	0.954
Percentile 90	1.673	1.563	1.776	1.363	1.378	1.782	2.026
Percentile 95	2.712	2.531	2.890	2.219	2.217	2.915	3.283
Percentile 99	7.277	6.669	7.671	5.855	5.911	7.700	8.808
Percentile 99.9	29.592	24.313	29.368	22.549	22.119	27.752	34.646

Table 62. Simulation results for exposure to manganese via tap water (µg/d) using 2010 concentration data and intake from DW16 2011 for children under 16 by body weight group.

Intake	DW16 2011 <17 kg	DW16 2011 17-26 kg	DW16 2011 26-41 kg	DW16 2011 ≥41 kg
Mean	0.642	0.734	0.831	1.025
Median	0.253	0.324	0.346	0.449
Upper quartile	0.609	0.714	0.794	1.012
Percentile 90	1.340	1.491	1.723	2.139
Percentile 95	2.195	2.399	2.803	3.448
Percentile 99	6.086	6.370	7.615	9.004
Percentile 99.9	21.670	22.500	27.832	29.939

Table 63. Simulation results for exposure to manganese via tap water on a body weight basis (μg $kg^{-1}d^{-1}$) using 2010 concentration data and intake from DW16 2011 for children under 16 by body weight group.

	Mean weight individual				Lightest individual			
Weight group, kg	< 17	17-26	26-41	≥ 41	< 17	17-26	26-41	≥ 41
Mean	0.05	0.03	0.03	0.02	0.13	0.04	0.03	0.02
Median	0.02	0.02	0.01	0.01	0.05	0.02	0.01	0.01
Upper quartile	0.05	0.03	0.02	0.02	0.12	0.04	0.03	0.02
Percentile 90	0.10	0.07	0.05	0.04	0.27	0.09	0.07	0.05
Percentile 95	0.17	0.11	0.09	0.06	0.44	0.14	0.11	0.08
Percentile 99	0.47	0.30	0.23	0.17	1.22	0.37	0.29	0.22
Percentile 99.9	1.67	1.05	0.84	0.56	4.34	1.32	1.07	0.73

7.7 Copper

7.7.1 Adults

The concentration distribution for copper was used during the analysis of the distributions for lead and iron because it had a similar shape but fewer values less than the LoD. A smaller set of exposure results is included here for comparison.

The results of simulating exposure to copper are shown in Table 64, and a typical distribution up to the 99.9th percentile (1 in 1000) is shown in Figure 22. Like those of those of iron and lead, the exposure distribution was highly skewed.

The 99.9th percentile predicted using the DWCS 2008 intake data (1.8 mg/d) exceeds the RNI for adults (1.2 mg/d; Buttriss, 2000) by about one-third and also exceeds the mean intake from other sources (1.06–1.39 mg/d; Bates *et al.*, 2011). However, because the distribution is highly skewed, both the mean (0.054 mg/d) and median (0.015 mg/d) are negligible compared with the RNI and the intake from other sources, and only 10% of exposures are predicted to exceed 10% of the RNI for copper via tap water. Thus, a tap water may occasionally make a significant contribution to the dietary requirement for copper, but it will generally be insignificant.

The PMTDI is 0.5 mg $kg^{-1}d^{-1}$ (FAO, 2012), which is equivalent to 20 mg/d for the lightest person in the NDNS 2001 data, more than 10 times the predicted 99.9th percentile and almost 400 times the mean exposure.

Table 64. Simulation results for exposure to copper in tap water using 2010 concentration data combined with all intake surveys containing adult respondents.

Intake	DWCS 1978 all	DWCS 1995 all	DWCS 2008 adults	NDNS 2001 adults
Chemical	Copper	Copper	Copper	Copper
Units	mg/d	mg/d	mg/d	mg/d
Year	2010	2010	2010	2010
Mean	0.040	0.048	0.054	0.046
Median	0.011	0.013	0.015	0.013
Upper quartile	0.034	0.040	0.045	0.038
Percentile 90	0.092	0.110	0.124	0.105
Percentile 95	0.163	0.198	0.220	0.187
Percentile 99	0.461	0.571	0.613	0.519
Percentile 99.9	1.293	1.592	1.798	1.610

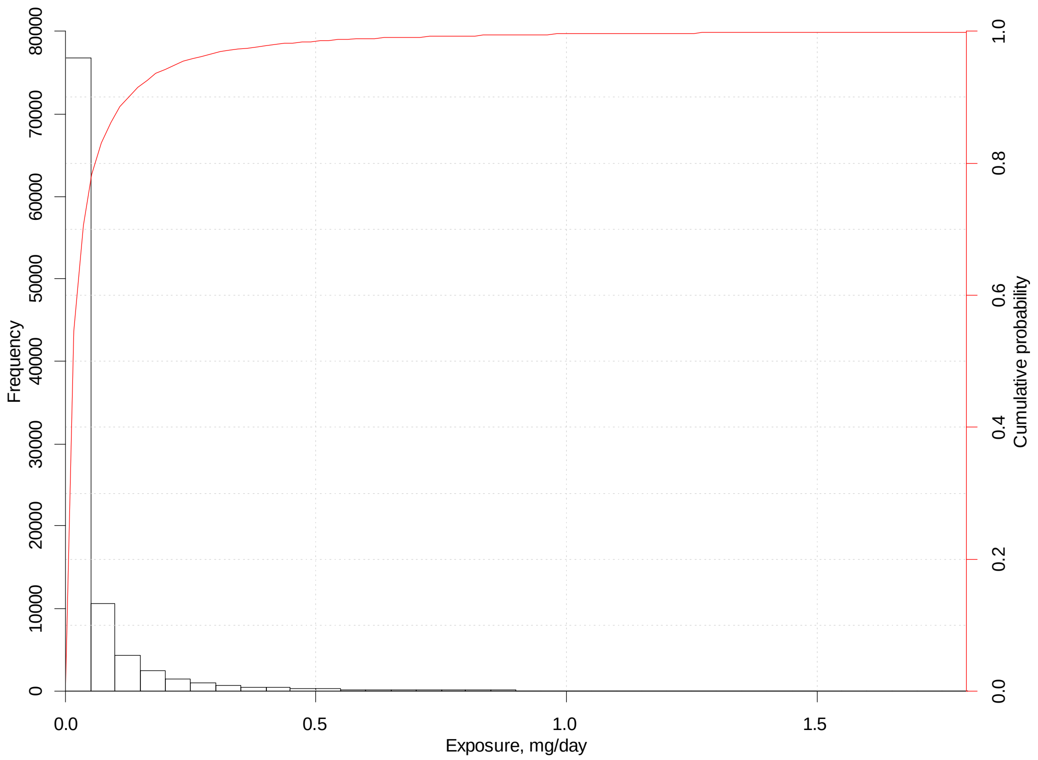


Figure 22. Simulation results (histogram and CDF) for exposure to copper in tap water using 2010 concentration data combined with DWCS 2008 intake data.

7.7.2 Children

The results of the simulation of exposure to copper for the whole sample and the sex and age groups are shown in Table 65, which includes the RNI for each age group (Buttriss, 2000). Ranges of RNIs are shown for the youngest and oldest groups, as explained in Section 7.2.2. As was the case for adults, the 99.9th percentile of the predicted exposure was slightly higher than the RNI in each group, though the mean was lower by a factor of 100. The mean total intake from food sources (Bates *et al.*, 2011) is about the same as the RNI. Therefore, drinking water normally makes an insignificant contribution to copper intake, but may occasionally be substantial.

The results for the four weight groups are shown in Table 66 and relative to the weights of the mean and minimum weight individuals in Table 67. The PMTDI is 500 μg $kg^{-1}d^{-1}$ (FAO, 2012) and the EVM Upper Safe Level is 160 μg $kg^{-1}d^{-1}$ (COT, 2003). The predicted 99.9th percentile for the <17 kg group calculated using the weight of the lightest individual was 128 μg $kg^{-1}d^{-1}$, four-fifths of the Upper Safe Level for total dietary intake, but only one-quarter of the PMTDI. The corresponding value for the other groups was 25–41 μg $kg^{-1}d^{-1}$. As noted above, the lightest individual in the <17 kg group is an extreme worst case who would be highly unlikely to have the highest intake, and the 99.9th percentile relates to a single exposure event, not persistent exposure. The predicted mean exposure was always less than 4 μg $kg^{-1}d^{-1}$.

Table 65. Simulation results for exposure to copper via tap water using 2010 concentration data and intake from DW16 2011 for under-16s by sex and age with RNIs (Buttriss, 2000).

Intake	DW16 2011 All	DW16 2011 Females	DW16 2011 Males	DW16 2011 Age 0-3	DW16 2011 Age 4-6	DW16 2011 Age 7-10	DW16 2011 Age 11-15
Units	mg/d	mg/d	mg/d	mg/d	mg/d	mg/d	mg/d
RNI				0.2–0.4	0.6	0.7	0.8–1.0
Mean	0.023	0.021	0.025	0.018	0.019	0.024	0.028
Median	0.006	0.005	0.006	0.004	0.005	0.006	0.007
Upper quartile	0.018	0.017	0.019	0.014	0.015	0.020	0.022
Percentile 90	0.051	0.047	0.055	0.042	0.043	0.055	0.064
Percentile 95	0.094	0.086	0.101	0.077	0.077	0.099	0.114
Percentile 99	0.267	0.250	0.295	0.228	0.222	0.290	0.330
Percentile 99.9	0.753	0.806	0.858	0.628	0.647	0.865	1.054

Table 66. Simulation results for exposure to copper via tap water (mg/d) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

Intake	DW16 2011 <17 kg	DW16 2011 17-26 kg	DW16 2011 26-41 kg	DW16 2011 ≥41 kg
Mean	0.018	0.021	0.024	0.030
Median	0.004	0.006	0.006	0.008
Upper quartile	0.014	0.017	0.019	0.024
Percentile 90	0.040	0.047	0.053	0.067
Percentile 95	0.074	0.084	0.097	0.123
Percentile 99	0.219	0.237	0.293	0.352
Percentile 99.9	0.640	0.695	0.879	1.021

Table 67. Simulation results for exposure to copper via tap water on a body weight basis (µg $kg^{-1}d^{-1}$) using 2010 concentration data and intake from DW16 2011 for under-16s by body weight group.

	Mean weight individual				Lightest individual			
Weight group, kg	< 17	17-26	26-41	≥ 41	< 17	17-26	26-41	≥ 41
Mean	1.39	0.98	0.73	0.56	3.61	1.24	0.92	0.73
Median	0.31	0.28	0.18	0.15	0.80	0.35	0.23	0.19
Upper quartile	1.08	0.79	0.58	0.45	2.81	1.00	0.73	0.58
Percentile 90	3.09	2.19	1.61	1.25	8.02	2.76	2.04	1.62
Percentile 95	5.71	3.91	2.94	2.30	14.83	4.94	3.73	2.98
Percentile 99	16.91	11.04	8.89	6.57	43.89	13.94	11.27	8.53
Percentile 99.9	49.42	32.39	26.66	19.07	128.26	40.88	33.81	24.73

7.8 Conclusions

This study has shown that probabilistic assessment of exposure via drinking water is readily feasible using the available data sources. It was possible to estimate information about the distribution of exposure, especially the probability of high values, that were not available from deterministic analyses using summary statistics. This approach can, therefore, give a more complete picture of both typical and extreme intakes.

The results of the simulations have shown that exposure to metals in tap water is highly variable. The 99.9th percentile exposure can be up to 45 times the mean and 200 times the median. It should be emphasised again that percentiles relate to the chance of individual daily exposures and the 99.9th percentile should not be interpreted as the regular daily exposure of 1 individual in 1000, which cannot be estimated reliably from the available data. The temporal variability in the concentrations measured at each location implies that the regular exposure would be much lower than the high percentile individual daily exposure.

The method of substitution for values less than the LoD had only moderate effects on the estimation of the mean and percentiles up to the 75th percentile for these determinands with many such values (e.g. lead and iron), and smaller effects on the statistics of the other determinands. The higher percentiles were unaffected in all cases. For simpler exposure assessments, substitution by either the LoD or LoD/2 would probably give acceptable accuracy.

Exposure to iron, lead, selenium and manganese predicted by the simulation appears to have decreased by about 40% between 2004 and 2010 due to falling concentrations in tap water. In particular, exposure to lead has decreased by 30–40%. For lead, this appears to be part a trend, halving previously decreased by 40% between 1994 and 2004. In contrast, the exposure to sodium appears to have increased slightly.

Comparing the predicted exposures with Reference Nutrient Intakes (for required nutrients) and Acceptable Daily Intakes or other recommended maximum intakes, we found for adults, using 2010 concentration data, that:

- For iron, selenium, sodium and manganese, the 99.9th percentile exposures were much less than the RNIs. In each case, the RNI is much lower than the ADI or similar upper limit.
- For copper, the 99.9th percentile exposure slightly exceeded the RNI and the intake from other sources, but the mean was very much smaller than the RNI, so tap water may occasionally make a significant contribution to the requirement for copper. The 99.9th percentile exposure was less than 10% of the PMTDI.
- For lead, the 99.9th percentile exposure was about 40% of the ADI (for lifetime exposure) in the worst case and the mean exposure was about 1% of the ADI. The ADI is currently being superseded by BMDL values. The mean exposure was about 5% of the $BMDL_{10}$ for nephrotoxity, which lay between the 99th and 99.9th percentiles of lead exposure in the worst case. Like the ADI, the BMDL relates to lifetime exposure, not acute effects, so the mean exposure is the most appropriate comparison.

For children under 16, using 2010 concentration data, we found that:

- For iron, the 99.9th percentile of predicted exposure was less than 5% of the RNI in all cases. In the worst case, the 99.9th percentile was less than 5% of the PMTDI.
- For selenium, the 99.9th percentile of predicted exposure was less than 20% of the RNI in most cases and less than 50% for the youngest age group. The mean exposure was less than 4% of the RNI. Thus, tap water may occasionally, but not persistently, be a significant contributor to nutrient intake. The 99.9th percentile is 0.01% of the Upper Safe Level in the worst case.
- For sodium, the 99.9th percentile of predicted exposure for the youngest group was about 30% of the RNI and the mean was about 3% of the RNI. The corresponding proportions for the other age groups were much smaller. The intake from other sources normally exceeds the RNI. Therefore, tap water may occasionally, but not frequently, be a significant contributor to sodium intake.
- For manganese, the 99.9th percentile of predicted exposure was less than 2% of the adequate daily intake for all groups aged 4 years and upwards. For the 0–3 years age group, the 99.9th percentile was less than 4% of the adequate daily intake for children aged over 6 months, but 7 times the adequate daily intake for babies up to 6 months. The data set is insufficient to allow this age group to be simulated separately, but it is possible that young babies may occasionally exceed the required nutrient intake. However, manganese has low acute toxicity, and no PMTDI has been set.
- For lead, the 99.9th percentile exposure was less than half of the ADI (for lifetime exposure) for most groups, but exceeded it slightly for the lightest group. The mean exposure was about 3% of the ADI in the worst case and about 1% of the ADI for the other groups. The mean exposure was less than 6% of the $BMDL_{01}$ (for long-term exposure) for developmental neurotoxicity in most cases and less than 20% in the worst case. The $BMDL_{01}$ was between the predicted 99th and

99.9th percentiles for most groups and between the 95th and 99th percentiles in the worst case. Thus the probability of persistently exceeding this level is relatively small.

8. Potential for application to Trihalomethanes (THMs)

THMs can be formed as a by-product of the use of chlorine to disinfect tap water and have been associated with potentially harmful effects in humans (WHO, 2004). They are monitored by the water companies and the Water Supply Regulations specify a maximum total concentration of 100 µg/l at consumers' taps for the four regulated THMs: chloroform, bromoform, bromodichloromethane and chlorodibromomethane (United Kingdom, 2000).

Simulation of the exposure of consumers to THMs through ingestion of drinking water would be possible using the methods that we have developed and applied here for metals. However, a complete exposure assessment would require the consideration of two other pathways of exposure – inhalation and dermal exposure – both of which are likely to occur principally while bathing and showering. These exposure pathways would be more complicated to represent than ingestion, which simply uses measured concentration data. Dermal exposure would also occur during washing and other tasks when water is handled. However, the duration and exposed skin area would normally be less than during showering and bathing, so the exposure is likely to be less significant. There could also be some inhalation exposure whenever water was heated and vented into the home.

Estimating THM exposure via dermal penetration during bathing or showering would require an estimation of potential THM skin penetration per unit area, which could be used in combination with statistical data for the frequency and duration of baths and body size (skin surface area). If reasonable assumptions could be made about dermal penetration, a Monte Carlo simulation model could be constructed along the lines described above, in which penetration over a given bathing time and for a given frequency of occurrence could be represented as a function of concentration and skin penetration rate.

Estimation of exposure to THM vapour or aerosols during bathing or showering would require information on the frequency and duration of the activity, bathroom ventilation, chemical concentration in water, chemical exchange between the dissolved and vapour phases (which will be a function of Henry's law constant, water temperature and, possibly, droplet size distribution and mass transfer coefficients; see McLachlan *et al.*, 1990), along with air intake rate to the lungs (breathing rate). The construction of such a model would not be straightforward but would be feasible. Some study data exist to inform model construction (e.g. Jo *et al.*, 1990) and some attempts have already been made to estimate exposure deterministically (using reasonable worst case assumptions). A good review and starting point can be found in Rockett *et al*. (2010).

The main exposure pathways are illustrated in Figure 23. It is similar to many of the models reviewed by IGHRC (2010), such as the Contaminated Land Exposure Assessment (CLEA) model from the Environment Agency (EA, 2009). However, many of these models are intended only to estimate average daily exposure, so are not probabilistic.

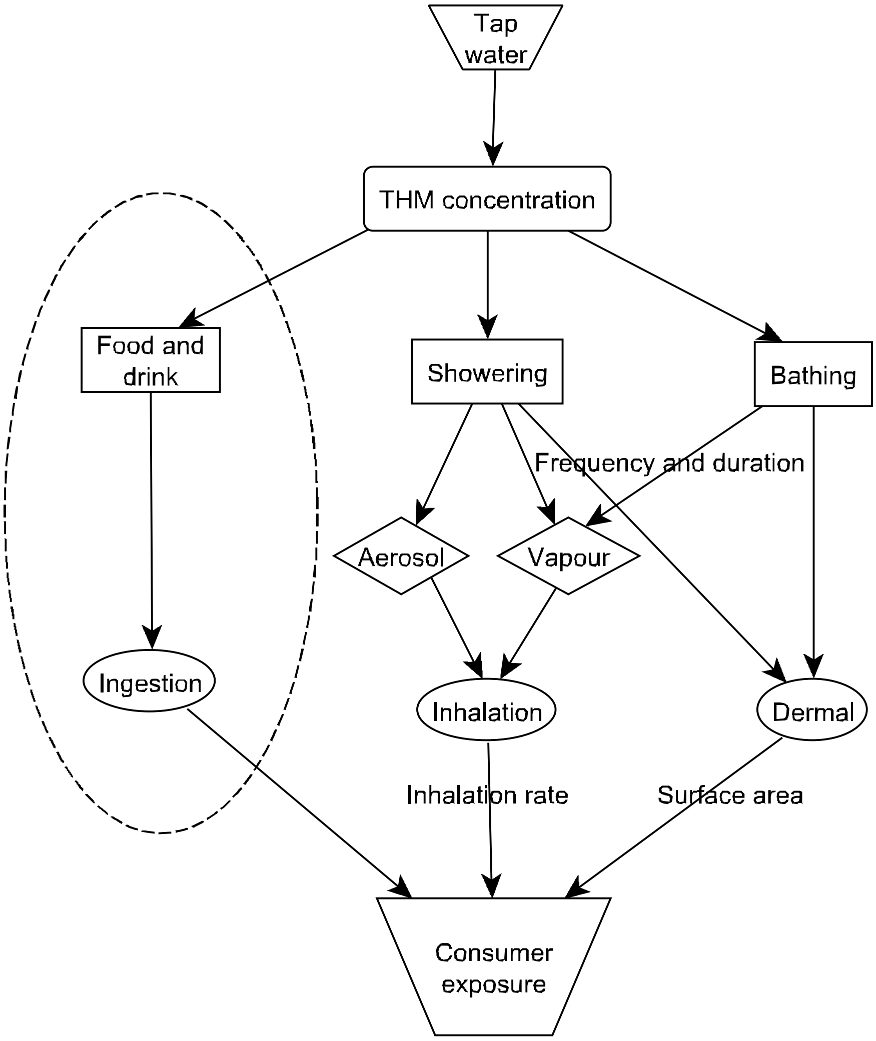


Figure 23. Potential main pathways of THM exposure. The oral ingestion pathway within the dashed oval could be modelled using the same methods as the other substances in this report; the others would require model development.

The total exposure (E_T) via non-ingestion routes while showering can be calculated as the sum of exposure via dermal penetration (E_D) and exposure via inhalation (E_I):

$$E_T = E_D + E_I \tag{1}$$

where all terms are in µg/day. Expanding this equation, we get

$$E_T = C_W J_D A D_S + (C_w I_A \varepsilon C_d + C_A I_A \varepsilon) D_s \tag{2}$$

where C_W is the concentration of the THM under consideration in the dissolved aqueous phase (µg/l), J_D is the dermal transfer rate (lm^{-2}min^{-1}), A is the skin area of a the person showering (m^2), D_S is the daily showering duration (min), I_A is the air inhalation rate (l/min), ε is a dimensionless lung absorption efficiency term (fraction of inhaled THM absorbed into the bloodstream), C_d is the concentration of liquid (aerosol) droplets in the air (l/l) and C_A is the concentration of THM in the air phase (μg/l). This is a simplification, assuming, in particular, the concentrations in air and water are constant for the duration of the bath or shower. A reasonable worst case would be to assume that the concentration in the water is not reduced significantly by volatilization, but that the THMs in the vapour phase are in equilibrium with the dissolved aqueous phase throughout. Note that the dermal transfer rate will depend on the partitioning properties of the chemical under consideration, which may be temperature dependent; for example there is evidence in the case of chloroform that it increases with temperature (Gordon *et al.*, 1998).

The air-water partition coefficient (K_{AW}), which is otherwise known as the dimensionless Henry's law constant, can be defined as

$$K_{AW} = C_A / C_W \tag{3}$$

so we can substitute for C_A in (2) to get

$$E_T = C_W J_D A D_S + (C_w I_A \varepsilon C_d + K_{AW} C_W I_A \varepsilon) D_s \tag{4}$$

which simplifies to

$$E_T = C_W D_S (A J_D + I_A \varepsilon C_d + I_A \varepsilon \mathrm{K_{AW}}) \tag{5}$$

K_{AW} is temperature dependent and this can be described as follows:

$$K_{AW}(T_e) = K_{AW}(T_r) \exp\left(\frac{\Delta U_{AW}}{R}\left(\frac{1}{T_R} - \frac{1}{T_E}\right)\right) \tag{6}$$

where T_e is the environmental (shower water) temperature (K), T_r is the reference temperature at which K_{AW} was derived (K), R is the gas constant and ΔU_{AW} is the enthalpy of phase change (J/mol).

Several of the terms in equations 5 and 6 are variable and could be described by probability density functions. These could be used in a Monte Carlo simulation, in the same manner as the concentration and intake were used for oral intake. At each iteration of the simulation, a random value for each uncertain quantity would be sampled from its distribution to calculate a single value for E_T. This would be repeated for a large number of iterations to generate a distribution of exposure values.

If the system were not in steady state, the concentration of THMs in the air would change over time. In the case of bathing, this would arise because the THMs were being volatilized from a relatively small surface area into a large volume of air. Showering is further complicated by the constant introduction of THMs into the system via flowing water. In both cases, the air space might be ventilated by windows or extractor fans. The extent to which the water and air are in thermodynamic equilibrium is also uncertain. It is likely that equilibrium would not be achieved, because of limitations in the volatilization of THMs imposed by resistance to intermedia diffusion (e.g. Whelan *et al.*, 2009), but this may be difficult to establish from existing data. Representing dynamic changes to air concentration would impose additional complexity on the exposure model, so it may be simpler initially to assume that equilibrium would be achieved. This assumption is conservative: it would probably overestimate the concentration in the air.

Deterministic models similar to this were used by Rockett *et al*. (2010) to obtain single estimates of exposure during bathing and showering via dermal absorption and inhalation. Most of the terms in the models were linear, but there was a nonlinear effect of shower duration, because they did not assume that the system was in equilibrium throughout. There was also a nonlinear, but relatively small, response to water temperature.

The key requirement in a probabilistic assessment would be the availability of suitable data. Concentration data for the four regulated THMs, as for the inorganic substances considered by this report, should be available from routine monitoring. The intake data used here would permit a straightforward assessment of oral exposure, though it may overestimate slightly due to volatilization of the THMs, particularly from hot liquids.

8.1 Example: chloroform

A partial example of how a probabilistic might be performed is presented here. Please note that this is intended to be illustrative of the method only and does not represent a fully defensible assessment.

Most of the earlier calculations by Rockett *et al*. (2010) were performed for each THM using the UK maximum permitted concentration for total THMs (100 μg/l) or the WHO guideline (300 μg/l) to represent a worst case (Rockett *et al.*, 2010), but this is extremely unlike to occur. Some calculations were also performed using concentrations derived from data on tap water: for example, the mean concentration of chloroform in tap water in England and Wales in 2000 was 12.5 μg/l.

A probabilistic assessment of exposure to chloroform requires a distribution for its concentration in tap water. Using the mean concentration given above (12.5 μg/l) and taking the maximum permitted total concentration of THMs (100 μg/l) as the 99.9th percentile, it is possible to obtain the parameters for a hypothetical lognormal distribution for concentration: $\mu = 2.230$, $\sigma = 0.768$ (Figure 24).

We simulated oral exposure by using this distribution with the NDNS 2001 tap water intake distribution (Table 68). The simulated mean intake of chloroform was 14 μg/d, the 99th percentile was 77 μg/d and the 99.9th percentile was 161 μg/d. The TDI for chloroform is 15 μg kg^{-1}d^{-1}, or 900 μg/d for a 60 kg adult, so the probability of exceeding this level from tap water intake, based on these assumptions, is negligible. The exposure is probably lower now, as the mean total THM concentration fell from 40 μg/l in 2000 to 25 μg/l in 2006 (Rockett *et al.*, 2010).

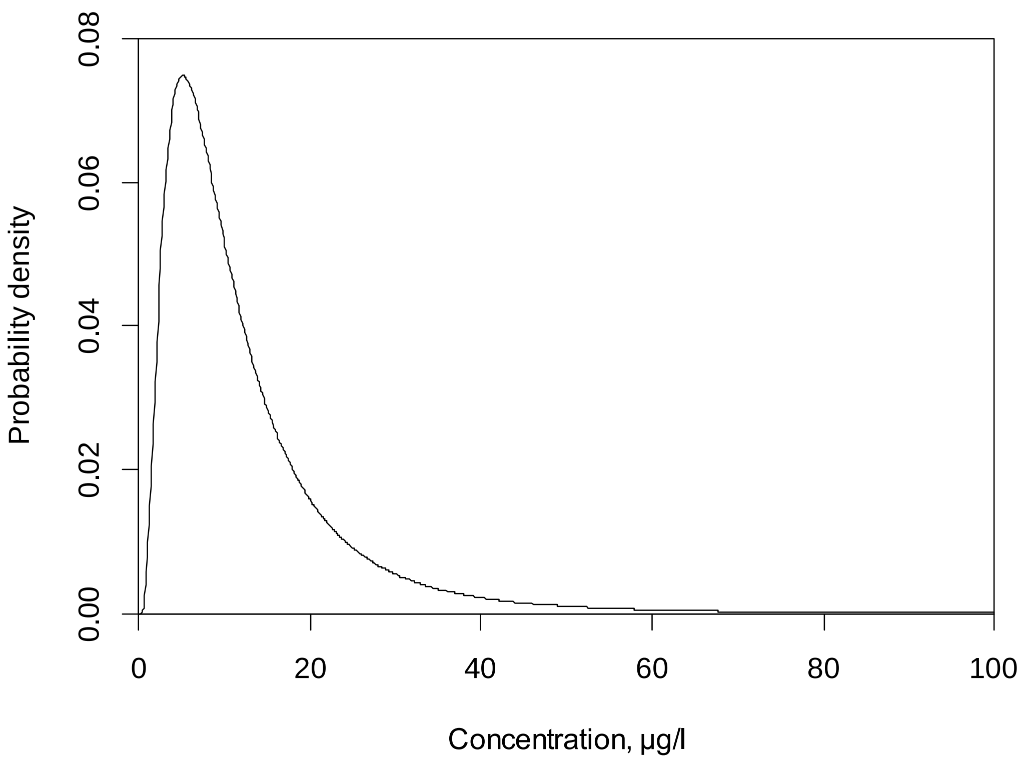


Figure 24. Hypothetical PDF for chloroform concentration in tap water, assuming mean = 12.5 µg/l and 99.9th percentile = 100 µg/l

Table 68. Simulation results for oral exposure to chloroform in tap water using the hypothetical concentration distribution for 2000 combined with a tap water intake distribution from NDNS 2001.

Statistic	Value, µg/d
Mean	13.79
Median	8.99
Upper quartile	16.95
Percentile 90	29.70
Percentile 95	41.60
Percentile 99	76.94
Percentile 99.9	160.98

Rockett *et al*. (2010) calculated that bathing for 30 minutes at a concentration of 12.5 µg/l resulted in an absorbed dose of chloroform via dermal exposure for a typical adult of 20.9 µg and that the same duration at 100 µg/l produced an absorbed dose of 167 µg. If all other variables in the model were kept constant, the relationship between concentration and absorbed dose was linear, so the hypothetical concentration distribution described above would be transformed to a lognormal distribution for dose with a mean of 20.9 µg and 99.9th percentile of 167 µg. They assumed a tolerable daily dose for chloroform of 57 μg $kg^{-1}d^{-1}$, or 4500 μg/d for a 60 kg adult, so this mean exposure was less than 0.5% of the tolerable daily dose and the hypothetical 99.9th percentile was less than 4% of the tolerable daily dose.

A complete probabilistic assessment would require distributions for other variables, including duration (Rockett *et al*. (2010) reported that 91% of individuals surveyed stated that they bathed for 30 minutes/day or less), temperature (assumed to be 35°C), wetted surface area and body mass. Clearly, the 9% of individuals who spend more time bathing would tend to have higher exposures, and their model had a linear relationship between absorbed dose and duration. However, it is possible that the internal body burden (at least in the epidermis) would approach equilibrium with the concentration in the water, producing a diminishing marginal response. Higher temperatures could slightly reduce dermal exposure by reducing the concentration of chloroform in the water as a consequence of higher volatilization rates, but increase exposure via inhalation. Conversely, there is evidence that the dermal penetration rate increases substantially with temperature (Gordon *et al.*, 1998). There would be some negative correlation between specific exposure (μg/kg) and body mass, because surface area decreases relative to mass as mass increases.

Rockett *et al.* (2010) found four estimates of the dermal transfer rate (J_D) for chloroform, some based on experiments, with the largest and smallest differing by a factor of 20. Rather than attempting to infer a distribution from these for simulation, it would be better to carry out simulations using each value, or the extremes. The estimates above used the largest estimated value.

Similar comparisons can be made for inhalation exposure during bathing and showering. For bathing, Rockett et al. (2010) found that the inhalation exposure was 51 μg/d at 12.5 μg/l and 404 μg/d at 100 μg/l (the 99th percentile of the hypothetical distribution). The latter figure is about 45% of the tolerable daily dose via this route. For a 15 minute shower the corresponding exposures were 62 μg at 12.5 μg/l and 495 μg (55% of the tolerable daily dose) at 100 μg/l. Note that all of these values, including the tolerable daily dose, represent the quantity inhaled and do not consider absorption of the chemical into the bloodstream.

As for dermal exposure, a probabilistic assessment of inhalation exposure would require distributions for several important variables, including duration, flow rate, air volume, breathing rate and temperature. The estimated exposure levels are greater relative to the tolerable dose than for the other routes, so there would be a greater chance of exceeding the tolerable dose if uncertainty in all the variables was included. To move from the inhaled dose to the absorbed dose, an estimate, with uncertainty, of the lung efficiency coefficient (ε) would also be needed.

As an illustration of the effect of including other uncertainties in the assessment of inhalation exposure while bathing, we assumed a normal distribution for duration, with a mean μ = 20 minutes and a standard deviation σ = 7.69, to give a 90th percentile of approximately 30 minutes, as reported by Rockett *et al*. As the normal distribution is not bounded below and could produce negative values, we truncated it at 0 (all negative values were set to 0). This distribution was used with the hypothetical distribution for chloroform concentration, while keeping all the other factors constant, in a simulation of 100,000 iterations. As expected, because the mean duration was shorter than 30 minutes, the mean simulated inhaled dose was reduced, to 34 µg/day. The 99.9th percentile dose was 313 µg/d (Table 69).

Table 69. Simulation results for inhalation exposure to chloroform in tap water during bathing based on the model used by Rockett *et al*. (2010) using the hypothetical concentration distribution for 2000 (see Figure 24) combined with an assumed normal distribution for daily bathing duration.

Statistic	Value, µg/d
Mean	34
Median	23
Upper quartile	42
Percentile 90	71
Percentile 95	96
Percentile 99	168
Percentile 99.9	313

The results obtained by Rockett *et al*. (2010) for bromoform and chlorodibromomethane were very similar to those for chloroform. However, for bromodichloromethane they found that the inhalation exposure during either bathing or showering, assuming a concentration of 100 µg/l, could exceed the tolerable daily dose. Therefore, it would be important to establish the true distributions of concentration and other key variables in this case.

9. Glossary

Censored data

Censoring occurs when the value of a measurement is only partially known. Common types include

- *Left censoring*: a measurement lies below a particular value (e.g. a limit of detection).
- *Right censoring*: a measurement lies above a particular value (e.g. the maximum of the scale on an instrument)
- *Interval censoring*: a measurement lies between two values (e.g. an instrument with fixed resolution).

Convolution product

The convolution product of two functions is an integral that expresses the degree of overlap of one function as it is shifted over the other. In the case of PDFs of two random variables, the convolution product gives the PDF of the product of the variables. For continuous random variables X and Y, with PDFs f_x and f_y, the product is

$$f_{XY}(u) = \int_{-\infty}^{\infty} f_X(z) f_Y(u-z)\,dz$$

The equivalent for discrete random variables is

$$f_{XY}(n) = \sum_{m=-\infty}^{m=\infty} f_X(m) f_Y(n-m)$$

Cumulative distribution function (CDF)

The CDF gives the probability that a random variable is less than or equal to a given value. It is the integral of the PDF (q.v.)

Exponential distribution

A continuous probability distribution specified by one parameter, λ (rate). Its CDF is

$$f_\lambda(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The PDF has it maximum at 0 and decreases monotonically, tending to 0 at ∞. The mean is $1/\lambda$.

Gamma distribution

A continuous probability distribution specified by two parameters, k (shape) and β (rate), or k (shape) and θ (scale), where $\theta = 1/\beta$. The PDF and CDF both depend on the Gamma function, and cannot be expressed using elementary functions. The PDF and CDF are zero for $x < 0$. When $k = 1$, it is identical to the exponential distribution. When

$k > 1$, the PDF is 0 at 0 and is positively skewed (the tail on the right is longer than the tail on the left), with the skew decreasing as k increases.

Likelihood (and log likelihood)

The likelihood function of a parameter θ of a distribution with PDF f_θ given an observation x is

$$\mathcal{L}(\theta|x) = p_\theta(x)$$

Informally, it gives the relative probability of different values of θ given the observation. Given a set of observations $\{x_i\}$, the likelihood is

$$\mathcal{L}(\theta|\{x\}) = \prod_i \mathcal{L}(\theta|x_i)$$

It is usually more convenient to work with the (natural) logarithm of the likelihood, which transforms the product above into a sum.

Lognormal distribution

The lognormal distribution is the distribution of a continuous variable whose natural logarithm is normally distributed. It is specified by the parameters of the underlying normal distribution, μ (mean) and σ (standard deviation). The PDF tends to 0 at 0 and ∞, and is positively skewed.

Normal distribution (Gaussian distribution)

A continuous probability distribution specified by two parameters, μ (mean) and σ (standard deviation). The PDF is symmetric, not skewed, and tends to 0 at $\pm\infty$, forming the shape known in the USA as a bell curve.

Maximum likelihood estimation (MLE)

MLE is a method of estimating the parameters of a statistical model given a set of observations by maximising the log-likelihood of the parameters given the observed data. It provides a unified approach to estimation, which is well defined for many distributions. In most simple cases it is equivalent to least squares regression (LSR), but generalises to problems (such as estimating the parameters of distributions), where LSR is inappropriate.

Positive skew

A positively skewed distribution has more of its weight below the mean than above it. Typically, the peak (mode) appears to be toward the left, with a long tail on the right, though this is not always the case. The distributions of chemical concentrations seen in this report are all strongly positively skewed and most of the intake distributions are slightly positively skewed.

Probability density function (PDF)

The PDF gives the relative probabilities of a random variable taking different values. For a continuous distribution (all the distributions used in this report), the probability of the variable taking a value within a given interval is the integral of the PDF (i.e. the area under the curve) within that interval. The PDF is the derivative of the CDF (q.v.)

Quantile and percentile

Quantiles represent proportions of the population arranged in order of magnitude. For example the 0.5 quantile is the middle value in the population and is known as the median. The 0.25 and 0.75 quantiles are known as the lower and upper quartile respectively. The most commonly used quantiles are the percentiles; for example the 25^{th} and 75^{th} percentiles are the 0.25 and 0.75 quantiles. When estimating a quantile from discrete samples the quantile may fall between two points (e.g. in the case of the median of a sample of 10 points, where it will lie between the fifth and sixth values), in which case its value is usually estimated by interpolation.

Quantile-quantile (Q-Q) plot

A Q-Q plot is a graphical method of comparing two distributions. A sequence of quantiles is chosen (for example the integer percentiles) and the value of each quantile from the second distribution is plotted against the value of the corresponding quantile from the first distribution. It is commonly used to compare a theoretical distribution with the data to which it was fitted. If the distributions are identical, the points will lie on a 1:1 straight line.

Weibull distribution

A continuous probability distribution specified by two parameters, k (shape) and λ (scale). The PDF is

$$f_{\lambda,k}(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{(-x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The PDF changes dramatically according to the value of k, but the case that is relevant for the data sets considered here is when $k > 1$: the PDF is 0 at 0 and positively skewed for small values of k (≤ 3.5).

10. References

Accent (2008). National tap water consumption Study. DWI 70/2/217 Phase Two Final Report, Accent, London for the Drinking Water Inspectorate, London

Anderson, T. W.; Darling, D.A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23, 193–212.

Armitage, P.; Berry, G.; Matthews, J.N.S. (2001). Statistical methods in medical research (4th edition). Blackwell Science, Oxford, 432 pp.

Auer, M. (2012). MonteCarlito 1.10 --- Free Excel Tool for Monte Carlo Simulation (accessed 10/06/2011). http://www.montecarlito.com/

Bates, B.; Lennox, A.; Bates, C.; Swan, G. (2011). National Diet and Nutrition Survey: Headline results from Years 1 and 2 (combined) of the rolling programme 2008/9 - 2009/10.. Department of Health. Available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsStatistics/DH_128166 (accessed June 2012),

Burmaster, D. E. (1998). Lognormal distributions for total water intake and tap water intake by pregnant and lactating women in the United States. *Risk Analysis*, 18, 215-220.

Buttriss, J. (2000). Nutrient requirements and optimisation of intakes. *British Medical Bulletin*, 56, 18-33.

COT (2003). Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment: 2000 total diet study of twelve elements (TOX/2003/39). Available at: http://cot.food.gov.uk/pdfs/TOX-2003-39.PDF,

CRAN (2012). The Comprehensive R Network. http://cran.r-project.org/

Chernobai, A.; Rachev, S.; Fabozzi, F. (2005). Composite goodness-of-fit tests for left-truncated loss samples. Technical Report, University of California, Santa Barbara

Crawley, M. J. (2007). The R book. John Wiley & Sons, Chichester, 942 pp.

DH (2009). UK–WHO growth chart 0–4 years. Department of Health, Royal College of Paediatrics and Child Health, World Health Organization. Available at http://www.rcpch.ac.uk/,

DWI (2011). Drinking water2010: public water supplies in the Central region of England. Drinking Water Inspectorate, London,

Defra (2010). Family Food 2010: Nutrient Intakes. Available at: http://www.defra.gov.uk/statistics/foodfarm/food/familyfood/,

EA (2009). Using science to create a better place: updated technical background to the CLEA model. Scince Report SC050021/SR1, Environment Agency, Bristol

EFSA (2010). Management of left-censored data in dietary exposure assessment of chemical substances. European Food Safety Authority, Parma, Italy

EFSA (2012). Lead dietary exposure in the European population. European Food Safety Authority, Parma, Italy,

Ershow, A. G.; Brown, L. M.; Cantor, K.P. (1993). Intake of Tap Water and Total Water by Pregnant and Lactating Women. *American Journal of Public Health*, 81, 328-334.

FAO (2012). Food safety and quality: Chemical risks and JECFA. Data from http://www.fao.org/food/food-safety-quality/scientific-advice/jecfa/en/, accessed May 2012

FSA (2002). Expert group on vitamins and minerals : review of manganese - revised version. EVM/99/22.REVISEDAUG2002. Food Standards Agency, London. Available at: http://www.food.gov.uk/multimedia/pdfs/evm9922p.pdf,

Gardner, M. (2011). Improving the interpretation of 'less than' values in environmental monitoring. *Water and Environment Journal*, , (To appear).

Gofti-Laroche, L.; Potelon, J. L.; Da Silva, E.; Zmirou, D. (2001). Description of drinking water intake in French communities. *Rev. Epidemiol. Sante Publique.* , 49, 411-422.

Gordon, S. M.; Wallace, L. A.; Callahan, P. J.; Kenny, D. V.; Brinkman, M.C. (1998). Effect of water temperature on dermal exposure to chloroform. *Environmental Health Perspectives*, 106, 337-345.

Henderson, L.; Gregory, J.; Swan, G. (2002). National Diet and Nutrition Survey: Adults Aged 19 to 64 years. Volume 1: Types and Quantities of Foods Consumed. HMSO, London,

Hilbig, A.; Kersting, M.; Sichert-Hellert, W. (2002). Measured consumption of tap water in German infants and young children as background for potential health risk assessments: data of the DONALD study. *Food Additives and Contaminants*, 19, 829-836.

Hopkin, S. M.; Ellis, J.C. (1980). Drinking water consumption in Great-Britain. A survey of drinking habits with special reference to tap-water based beverages. Technical Report TR137, Water Research Centre, Medmenham

Hydes, O. D.; Jiggins, P. L.; Marsden, P. K.; Davies, A.; Quennell, S.; Jackson, C.R. (1996). Nitrate pesticides and lead 1991 to 1994. Drinking Water Inspectorate, London,

Hydes, O. D.; Jiggins, P. L.; Marsden, P. K.; Walls, R. M.; Jackson, C.R. (1997). Nitrate pesticides and lead 1995 and 1996. Drinking Water Inspectorate, London,

IGHRC (2004). Guidelines for good exposure assessment practice for human health effects of chemicals (cr10). MRC Institute of Environment and Health, University of Leicester, MRC Institute for Environment and Health, Leicester

IGHRC (2010). Current approaches to exposure modelling in UK Government departments and agencies (cr15). Institute of Environment and Health, Cranfield University,

Ipsos MORI (2012). Tap water drinking behaviour: as study of children 0-15. Report for Defra's Drinking Water Inspectorate (DWI), MEL Research, Birmingham

Jensen, F. V. (1996). An introduction to Bayesian networks. UCL Press, London, UK, pp.

Jo, W. K.; Weisel, C. P.; Lioy, P.J. (1990). Routes of chloroform exposure and body burden from showering with chlorinated tap water. *Risk Analysis*, 10, 575-580.

Kaplan, E. L.; Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Kullback, S.; Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22, 79-86.

Levallois, P.; Guevin, N.; Gingras, S.; Levesque, B.; Weber, J. P.; Letarte, R. (1998). New patterns of drinking-water consumption: results of a pilot study. *The Science of the Total Environment*, 209, 233-241.

MEL Research (1996). Tap water consumption in England and Wales: findings from the 1995 national survey. Research Report 9448/01, MEL Research, Birmingham

Marsden, P. K. (2010). Addendum to national tap water consumption report. DWI 70/2/217, Drinking Water Inspectorate, London

McCullagh, P.; Nelder, J. (1989). Generalized Linear Models. Chapman and Hall, London, 532 pp.

McLachlan, M.; Mackay, D.; Jones, P.H. (1990). A conceptual-model of organic-chemical volatilization at waterfalls. *Environmental Science & Technology*, 24, 252-257.

Mons, M. N.; Blokker, E. J. M.; van der Wielen, J. M. L.; Medema, G. J.; Sinclair, M. I.; Hulshof, K. F. A. M.; Dangendorf, F.; Hunter, P.R. (2005). Estimation of the consumption of cold tap water for microbiological risk assessment. Report from the microrisk project. Microbiological risk assessment: a scientific basis for managing drinking water safety from source to tap,

Mons, M. N.; van der Wielen, J. M. L.; Blokker, E. J. M.; Sinclair, M. I.; Hulshof, K. F. A. M.; Dangendorf, F.; Hunter, P. R.; Medema, G.J. (2007). Estimation of the consumption of cold tap water for microbiological risk assessment: an overview of studiesand statistical analysis of data. *Journal of Water and Health*, 0.5 Suppl 1, 151-170.

Palisade Corporation (2007). @RISK 5.0 risk analysis software. http://www.palisade.com/risk/

Parsons, D. J.; Orton, T. G.; De Souza, J.; Moore, A.; Jones, R.; Dodd, C.E.R. (2005). A comparison of three modelling approaches for quantitative risk assessment using the case study of Salmonella spp. in poultry meat. *International Journal of Food Microbiology*, 98, 35-51.

Rockett, L.; Ewence, A.; Gardner, M.; Young, W.; Rumsby, P. (2010). A review of the current knowledge on dermal and inhalation toxicity of regulated THMs. WRc Report to the DWI. WRc Ref: UC7907.06 DWI Ref: 70/2/230,

Roseberry, A. M.; Burmaster, D.E. (1992). Lognormal distributions for water intake by children and adults. *Risk Analysis*, 12, 99-104.

SACN (2003). Salt and health. Scientific Advisory Committee on Nutrition. The Stationery Office, London,

SCF (1993). Reports of the Scientific Committee on Food, 31st series. Nutrient and energy intakes for the European Community. Commission of the European Communities, Luxembourg,

SQLite Consortium (2012). SQLite (accessed 10/6/2012). http://www.sqlite.org/

Sichert-Hellert, W.; Kersting, M.; Manz, F. (2001). Fifteen year trends in water intake in German children and adolescents: Results of the DONALD Study. *Acta Paediatrica*, 90, 732-737.

USDA (2011). Drinking water intake in the U.S. What We Eat In America, NHANES 2005-2008. National Health and Nutrition Examination Survey (NHANES). Food Surveys Research Group Dietary Data Brief No.7, September 2011,

United Kingdom (2000). The Water Supply (Water Quality) Regulations 2000 (England and Wales). Statutory Instrument 2000 No. 3184,

Varela, J. R. (2011). Simulación 4.0 (accessed 09/05/2011). http://www.ucema.edu.ar/u/jvarela/index_eng.htm

WHO (1987). Report of the 30th Meeting of the Joint FAO/WHO Expert Committee on Food Additives. World Health Organization, Rome, June 2-11, 1986. Published: Geneva, 1987,

WHO (2000). Human exposure assessment. International Programme on Chemical Safety, Environmental Health Criteria 210, World Health Organization, Geneva

WHO (2001). Human vitamin and mineral requirements. Report of a joint FAO/WHO expert consultation Bangkok, Thailand. Food and Agriculture Organization of the United Nations and World Health Organization, Rome,

WHO (2004). Trihalomethanes in drinking-water. Background document for development of WHO guidelines for drinking-water quality . World Health Organization. WHO/SDE/WSH/03.04/64,

Westrell, T.; Andersson, Y.; Stenstrom, T.A. (2006). Drinking water consumption patterns in Sweden. *Journal of Water and Health*, 04.4, 511-522.

Whelan, M. J.; Sanders, D.; van Egmond, R. (2009). Effect of Aldrich humic acid on water & atmosphere transfer of decamethylcyclopentasiloxane. *Chemosphere*, 74, 1111-1116.

Probabilistic modelling for assessment of exposure via drinking water

Final Report of Project Defra WT1263 / DWI 70/2/273

Appendix A. Examples of implementations of simple exposure models using different tools

1. Introduction

Section 1.1 of the report gave an overview of different methods and tools for constructing an exposure model. This appendix gives a little more detail and shows the results from each method. The model used is for exposure as the product of intake and concentration. Intake has a gamma distribution with shape 3.29 and rate 2.54. Concentration has an exponential with mean 2, hence rate 0.5.

2. Monte-Carlo Simulation

2.1 Excel with Monte-Carlito

Monte Carlito (Auer, 2012) is a set of macros written in Visual Basic for Applications™ for Microsoft Excel™ released under the GNU General Public Licence. It relies on the normal functions of Excel to generate random numbers. A screenshot is shown in Figure 1. The parameters of the intake distribution are in cells B4 and B5. The formula to generate the distribution is in B12:

```
=GAMMAINV(RAND(), $B$4, 1/$B$5)
```

This is the standard method of generating random variates in Excel: the output from the uniform [0, 1] random number generator is passed to the inverse of the CDF for the distribution – in this case the gamma distribution. Note that the gamma distribution in Excel uses the shape and scale parameters, where the scale is the reciprocal of the rate (1/$B$5).

Similarly, the mean for the concentration distribution in in cell B8 and the formula to generate it is in C12. Excel does not explicitly provide the inverse CFD for the exponential distribution, but it can easily be constructed using the LN function or, as in this case, from the gamma distribution with shape = 1:

```
=GAMMAINV(RAND(), 1, $B$8)
```

The formula for the exposure is in D12:

```
=B12*C12
```

Monte Calito requires all the cells containing the inputs and outputs of the simulation to be placed in a single row in this way, with the first cell (A12) containing the required number of iterations – in this case 100,000. The minus sign is used to tell the macro to minimise the windows while simulating, which speeds up the run, because Excel does not have to update the display. The seven rows below row 12 show the summary statistics from the simulation, which are filled in when the macro is run.

The summary statistics did not include the percentiles, so the macro was modified to output the results of each iteration to successive rows in a second sheet and to calculate additional statistics, as shown in Table 1. Monte Carlito can also produce histograms of each of the variables using standard Excel graphics (Figure 2).

The advantages of Monte Carlito are its simplicity and the fact that it can be modified, as we did, to add missing features. However, it is limited by relying on the random number generator and distribution functions available in Excel and lacks more advanced features, such as specifying the correlations between variables. A run using 65,000 iterations took about 30 seconds.

	A	B	C	D
1	**Simple exposure simulation using Excel and MonteCarlito**			
2				
3	Intake - gamma distribution			
4	Shape	3.29		
5	Rate	2.54		
6				
7	Concentration - exponential distribution (implemented as gamma with shape=1)			
8	Mean	2		
9				
10				
11	N	Intake, l/d	Concentration, mg/l	Exposure, mg/d
12	***-100000***	0.6761	5.8493	3.9550
13	Mean	1.2979	1.9974	2.5987
14	Standard error	0.0023	0.0063	0.0105
15	Median	1.1660	1.3898	1.4913
16	Standard deviatic	0.7186	2.0006	3.3233
17	Variance	0.5164	4.0025	11.0444
18	Skewness	1.1441	2.0204	3.4129
19	Kurtosis	5.0756	9.1643	23.7812
20				

Figure 1. Simple exposure simulation in Excel with Monte Carlito

Table 1. Additional statistics from the Monte Carlito simulation

	Intake	Concentration	Exposure
Mean	1.297857	1.997434	2.598671
Std dev	0.718604	2.000617	3.323308
Percentile 1	0.214726	0.020115	0.018526
Percentile 2.5	0.297611	0.051758	0.047425
Percentile 5	0.384023	0.103118	0.09433
Percentile 10	0.508657	0.212997	0.197522
Percentile 25	0.77249	0.575241	0.565567
Percentile 50	1.16605	1.389863	1.491343
Percentile 75	1.67877	2.762896	3.345272
Percentile 90	2.264586	4.600101	6.25706
Percentile 95	2.665252	5.972837	8.871788
Percentile 97.5	3.040854	7.394328	11.66553
Percentile 99	3.52095	9.216878	15.89718
Max	7.63139	24.51418	65.77519

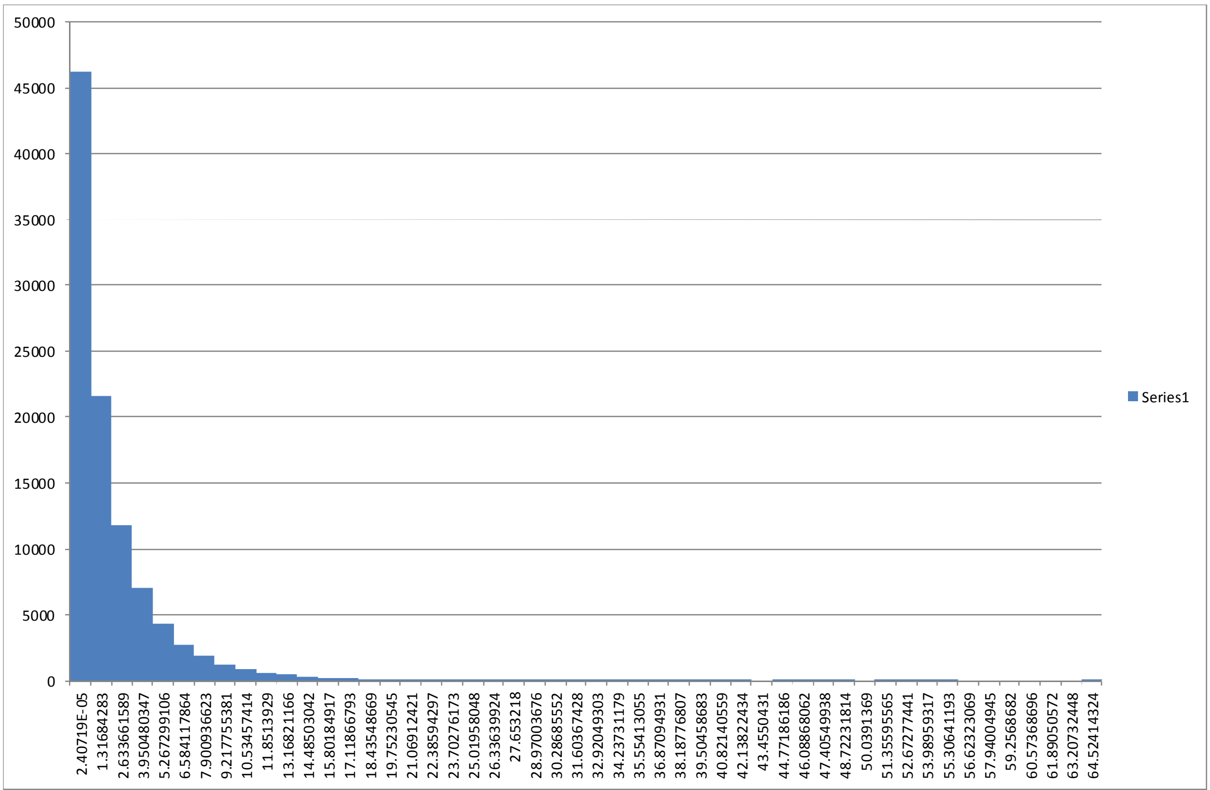


Figure 2. Exposure histogram from the Monte Carlito simulation

2.2 Excel with Simulación

Simulación (Varela, 2011) is an add-in for Excel, which is distributed as freeware in the form of a 600 KB Excel file. Once installed it adds commands to the Add Ins ribbon (Table 1).

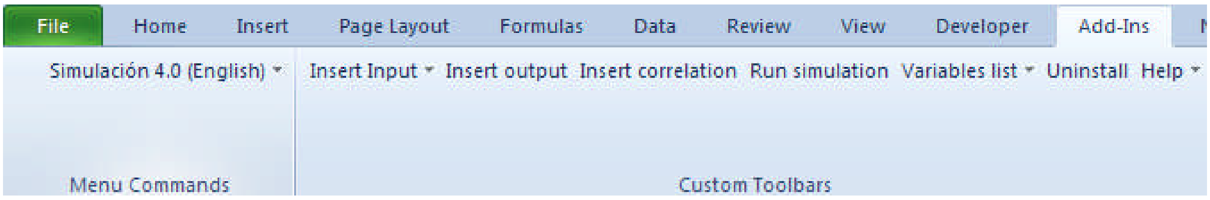


Figure 3. The Simulación menu on the Add-Ins ribbon in Excel

The model (Figure 4) was set up in a very similar way to Monte Carlito, except that the input distributions used functions provided by Simulación and inserted using the menu, so the intake (B12) and concentration (C12) were

```
=Simula_Gamma(0,B4,1/B5,"Intake")
=Simula_Exponencial(0,B8,"Concentration")
```

The output (D12) was also identified by use of a function:

```
=B12*C12+simula_output("Exposure")
```

	A	B	C	D
1	**Simple exposure simulation using Excel and Simulacion**			
2				
3	Intake - gamma distribution			
4	Shape	3.29		
5	Rate	2.54		
6				
7	Concentration - exponential distribution			
8	Mean	2		
9				
10				
11		Intake, l/d	Concentration, mg/l	Exposure, mg/d
12		1.2953	2.0000	2.5906
13				

Figure 4. Simple exposure simulation in Excel with Simulación

Although the same layout as before was used, this was not required by Simulación: any cells could be used as inputs and outputs, up to maxima of 150 inputs and 20 outputs.

The model was run by choosing *Run simulation* from the menu and setting the number of iterations, but it was limited to a maximum of 65,000 iterations. The results were presented in a separate Excel workbook, which contained pages for the raw results, summary statistics (Table 2), more detailed statistics and sensitivity analysis. The last two contained drop down lists to choose the variable to display. The detailed statistics page included a histogram for the selected variable (Figure 5), which was very small by default, but could be zoomed to make it more readable.

Unfortunately, there appeared to be an error in the gamma distribution for intake. With the parameters given, the mean should be 1.295 and the standard deviation 0.714. The mean was correct, but the standard deviation was much too large, producing an implausibly high maximum intake, with a consequent effect on the exposure. Further investigation would be needed to discover whether this was user error or a fault in the program.

Table 2. Summary statistics from the Simulación simulation

Name	Maximum	Minimum	Mean	Variance	Std. Dev.	Dev./Mean
Exposure	210.7916	0.0000	2.6085	41.8223	6.4670	247.92%
Concentration	21.2290	0.0001	1.9972	4.0128	2.0032	100.30%
Intake	29.2806	0.0000	1.2962	4.2350	2.0579	158.76%

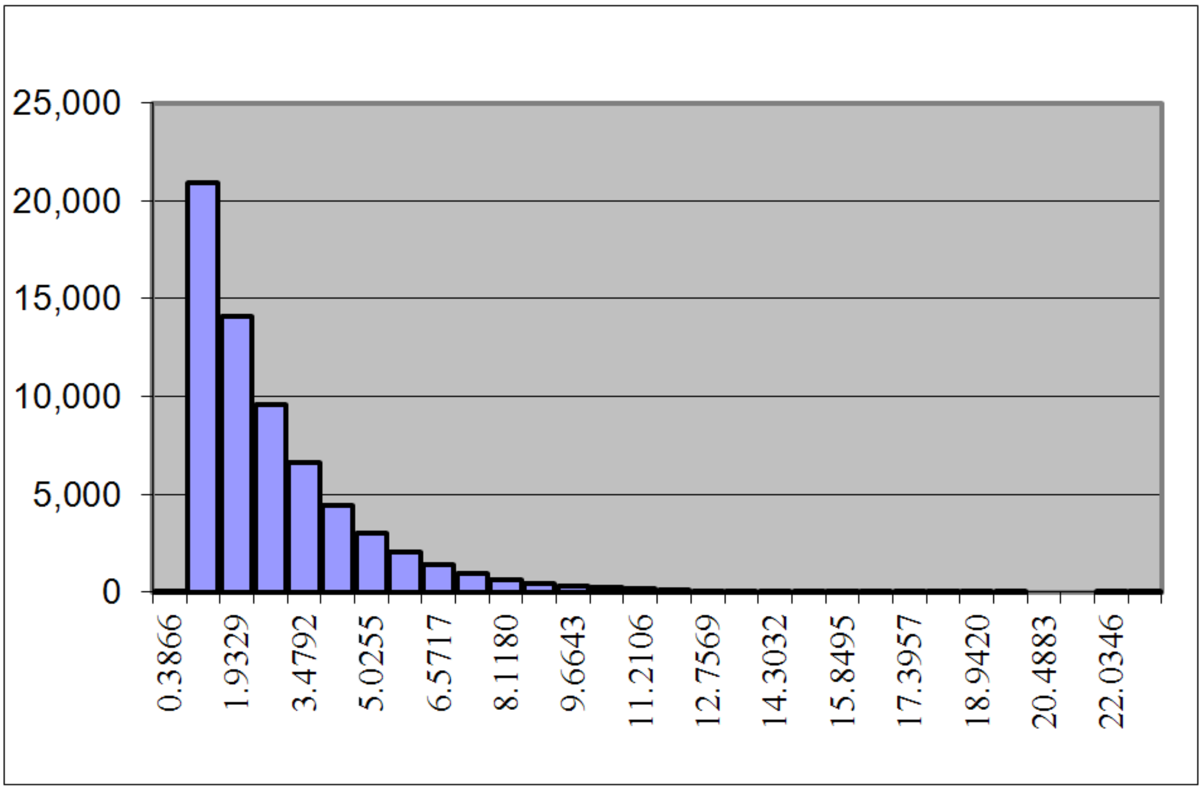


Figure 5. Histogram of concentration from the Simulación simulation

Two important features of Simulación compared with Monte-Carlito are the wider choice of distributions and the ability to specify correlations between input variables. For this simulation it was slower than Monte Carlito, taking about 70 seconds for 65,000 iterations, though we have found it to be quicker than Monte Carlito for other models.

2.3 Excel with @RISK

@RISK™ (Palisade Corporation, 2007) is a commercial add-in for Microsoft Excel produced by Palisade Corporation. It is a comparatively large package (~100 MB) and provides a wide range of features, many of which are not available in the free alternatives, including fitting distributions to data. When installed it adds its own ribbon to Excel (Figure 6).

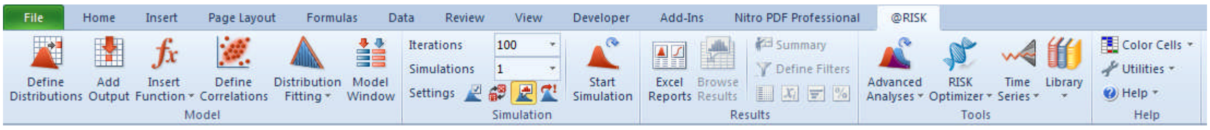


Figure 6. The @RISK ribbon in Excel

Setting up the model was almost identical to Simulación, by using the @RISK ribbon to insert functions for the input distributions

```
=RiskGamma(B4,1/B5,RiskShift(0),RiskStatic(1))
=RiskExpon(B8,RiskShift(0),RiskStatic(1))
```

and to collect the output

```
=RiskOutput()+B12*C12
```

It was run by for 100,000 iterations using the buttons on the ribbon.

The initial output shows a histogram and summary statistics for the selected cell (Figure 7). The statistics can be replaced by the raw simulation results, and a wide range of additional graphs, reports and analyses is available. The graphs are produced directly by @RISK, which offers more options than standard Excel graphics. Some of the summary statistics are shown in Table 3 for comparison with the other versions; the results are very similar to those from Monte Carlito.

Table 3. Selected summary statistics from the @RISK simulation

	Intake	Concentration	Exposure
Minimum	0.0216	1.862E-006	2.151E-006
Maximum	6.7737	35.131	53.402
Mean	1.2953	2.000	2.590
Median	1.1667	1.386	1.485
Std dev	0.7141	2.001	3.274
Skewness	1.1016	2.0258	3.2164
Kurtosis	4.8111	9.5485	20.5235
90^{th} percentile	2.2528	4.605	6.309
95^{th} percentile	2.6480	5.991	8.802
99^{th} percentile	3.5003	9.210	15.625

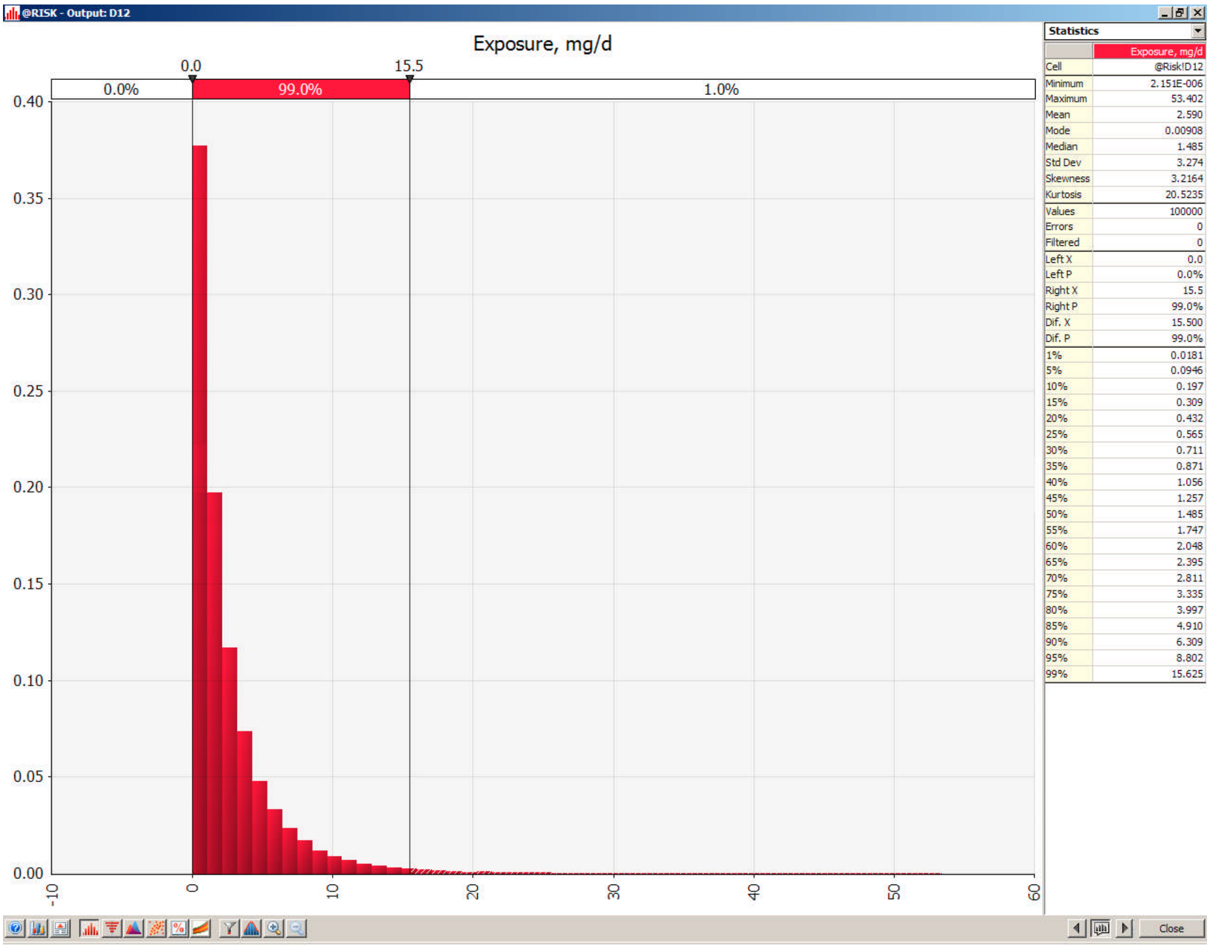


Figure 7. Histogram and summary statistics for concentration from the @RISK simulation

As expected from a well-established commercial product, @RISK has a much more extensive set of features and lacks many of the limitations of the free alternatives. It was also substantially quicker, taking only 6 seconds for 65,000 iterations.

2.4 Monte-Carlo Simulation in R

Working in R (Crawley, 2007) uses a completely different approach to Excel. It is a programming language, rather than an interactive tool for manipulating data. Commands can either be typed into a console window, in which case they are executed immediately, or saved in a text file and run as a complete 'script'.

For comparison with the Excel-based methods, a minimal script to create the same simulation as above and its output are shown in Figure 8, and the resulting histogram in Figure 9. Other than specifying the number of class breaks for the histogram, the default options for printing and graphics have been used; much better formatting is possible with additional programming. Summary statistics are show in Table 4; the results are similar to those from Monte Carlito and @RISK.

```
# Minimal simulation of exposure in R.
# Number of cases to simulate
N = 100000

# Parameters for intake distribution (l/d)
IntakeShape = 3.29
IntakeRate = 2.54

# Parameter for contamination distribution (mg/l)
ConcMean = 2

# Input distributions
concentration = rexp(N, 1/ConcMean)
intake = rgamma(N, IntakeShape, IntakeRate)

# Output: exposure
exposure = concentration * intake

# Print the results without formatting
print (c(mean(exposure), sd(exposure), min(exposure), max(exposure),
      quantile(exposure, c(0.9, 0.95, 0.99))))

# Plot histogram
hist(exposure, breaks=30)

--------------------------------------------------------------------------
# Output
2.607007e+00 3.292558e+00 2.157791e-05 6.154171e+01
         90%          95%          99%
6.299985e+00 8.810924e+00 1.587904e+01
```

Figure 8. Minimal simulation of exposure in R

Clearly, R is more of a tool for specialists than the Excel-based methods. It has many features (such as specifying correlations between variables) available as libraries written by the user community and is essentially limited only be the skill of the user. It was two orders of magnitude quicker to run: 6,500,000 iterations took 10 seconds (equivalent to 0.1 seconds for 65,000 iterations).

Table 4. Selected summary statistics from the simulation in R

	Intake	Concentration	Exposure
Minimum	0.03	0.00	0.00
Maximum	6.84	21.73	61.54
Mean	1.30	2.01	2.61
Median	1.17	1.39	1.50
Std Dev	0.72	2.01	3.29
Skewness	1.12	1.98	3.24
Kurtosis	4.93	8.67	20.82
90th percentile	2.26	4.62	6.30
95th percentile	2.66	6.01	8.81
99th percentile	3.52	9.21	15.88

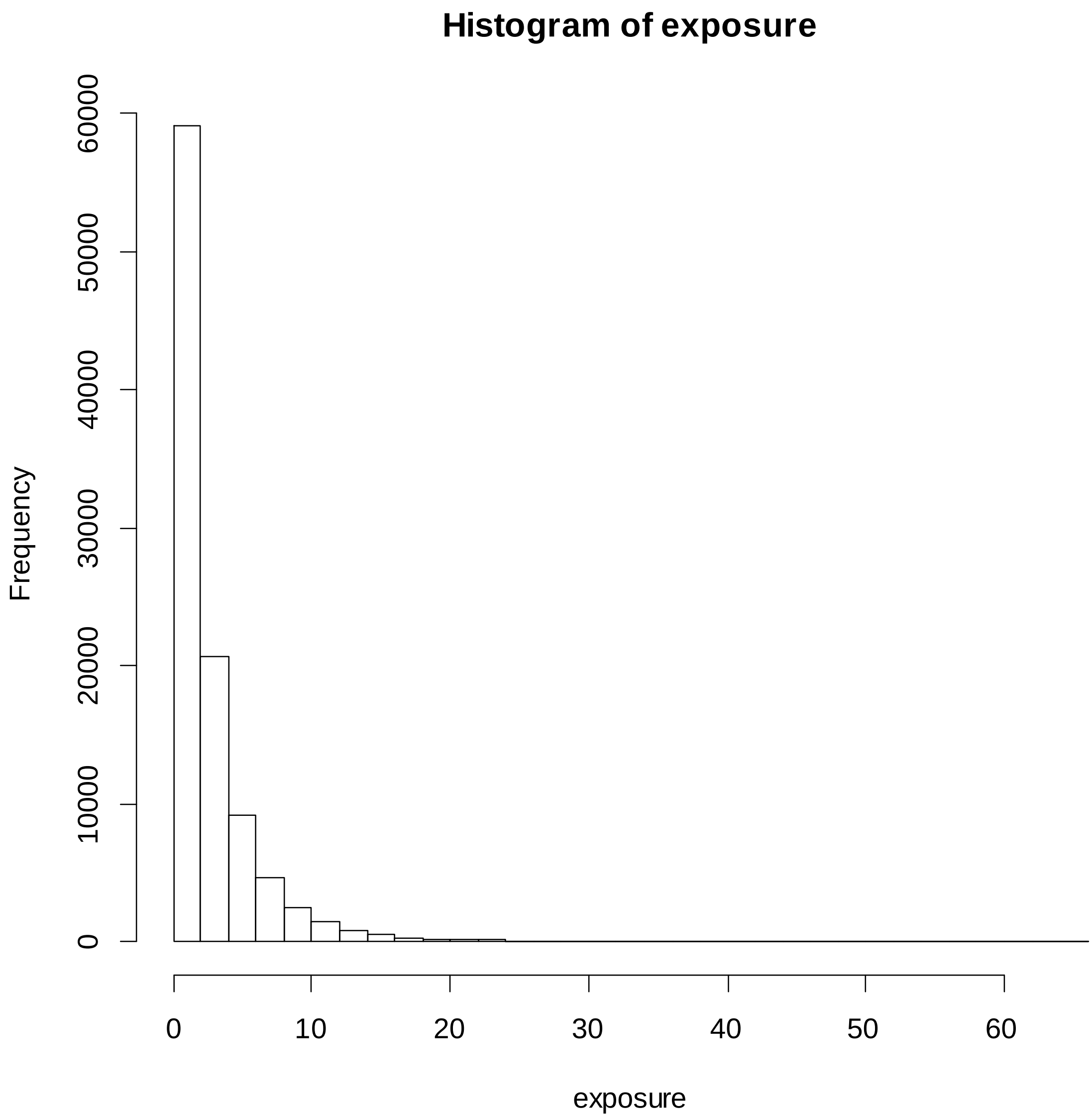


Figure 9. Histogram of exposure from the simulation in R

3. Bayesian models

3.1 Bayesian network in Netica

Bayesian networks (also known as Bayesian belief networks or causal probability networks) are a relatively new approach to probabilistic modelling (see e.g. Jensen, 1996). A simple Bayesian network version of the exposure model was constructed in Netica™ (Norsys Software Corporation, Vancouver, Canada), which provides a graphical user interface for constructing models. A model consists of several nodes (variables), joined by arrows representing causal relationships. The variables are normally quantified in the form of probability tables and conditional probability tables.

In this case, the model contained three nodes: two 'parent' nodes – Intake and Concentration – and one 'child' node – Exposure (Figure 10).

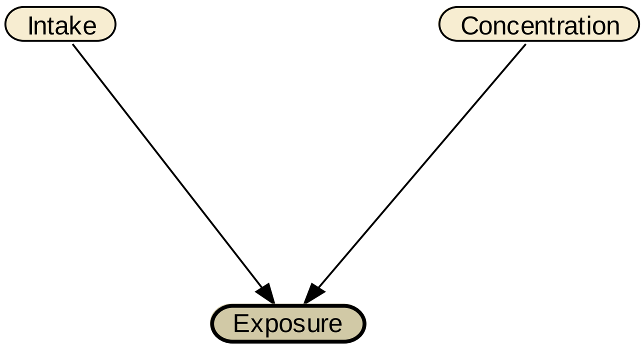


Figure 10. Simple network model for exposure, depending on intake and concentration.

The variables in Bayesian networks are normally specified by tables containing the probability distributions (or conditional probabilities for child nodes). However, in this case, equations were used to specify intake and concentration as distributions

```
p(Intake) = GammaDist(Intake, 3.29, 1/2.54)
p(Concentration)= ExponentialDist(Concentration, 1/2)
```

and exposure as their product

```
Exposure (Intake, Concentration) = Intake * Concentration
```

In order to use the model, the variables had to be converted to discrete versions by dividing their ranges into small steps, then sampling the distributions to generate probability tables. It was then 'compiled' to calculate the results for Exposure. These steps were largely automatic and performed quickly by the software. The result was histograms for the three variables (Figure 11), which also showed the mean and standard error. None of the other summary statistics was readily available.

Typical uses of a Bayesian network would be to fix the value of one or more inputs to see the effect on the output, or conversely to fix the output to see the most likely combination of inputs to give that result (known as inference). These features are of limited use for this model. Changing the parameters of the model for different populations or chemicals required the tables to be re-generated.

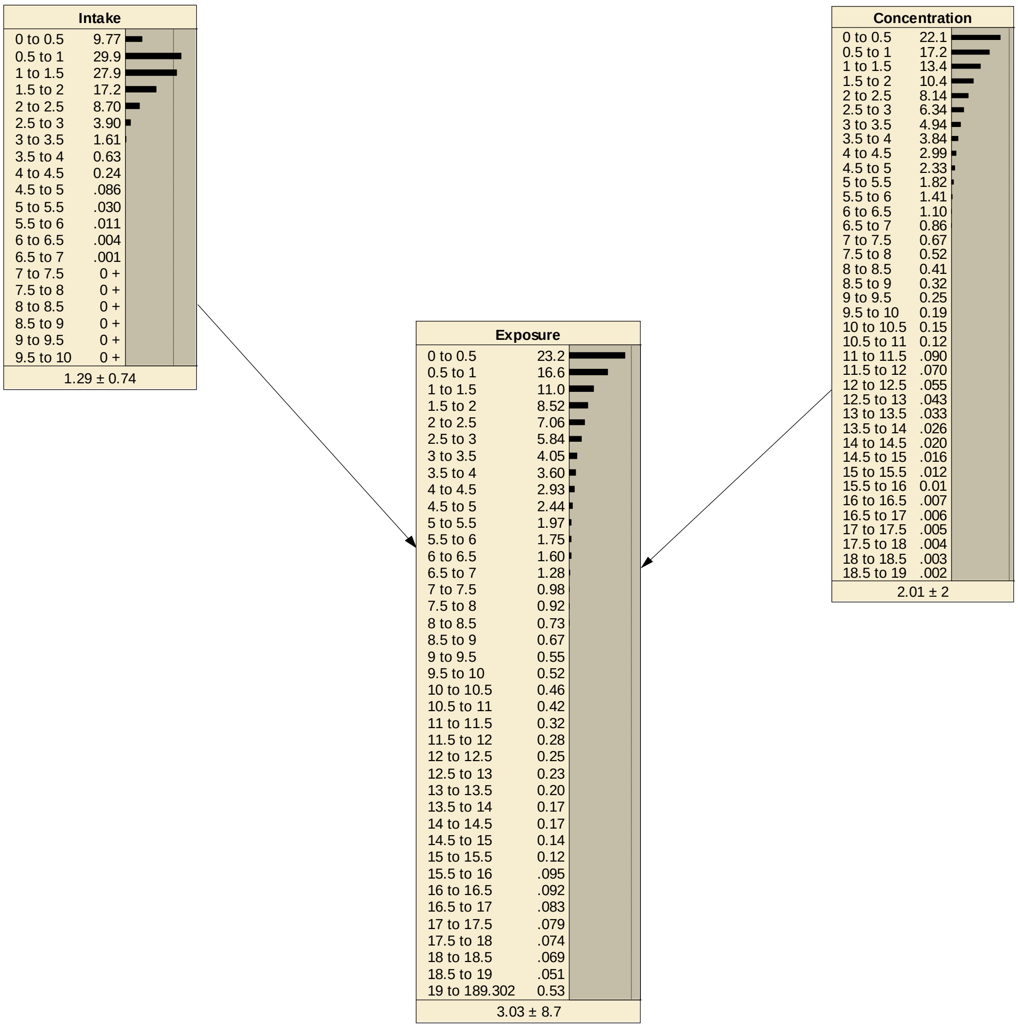


Figure 11. Bayesian network model for exposure expanded to show distributions with mean and standard error below

3.2 Markov-Chain Monte Carlo

Markov Chain Monte-Carlo (MCMC) methods are a class of algorithms for generating probability distributions. One application of them has been in a freely available software package called BUGS (Bayes Using Gibbs Sampler) and now OpenBUGS (http://www.openbugs.info/w/). The program can be used to construct Bayesian network models, but solves them using simulation instead of direct inference. The advantage of this approach is that continuous distributions can be used without discretization and models (such as the product of concentration and intake) are calculated directly.

The same model as above was constructed using OpenBUGS (Figure 12), with the details specified as three equations:

```
model{
    Concentration ~ dexp(0.5)
    Intake ~ dgamma(3.29, 2.54)
    Exposure <- Intake * Concentration
}
```

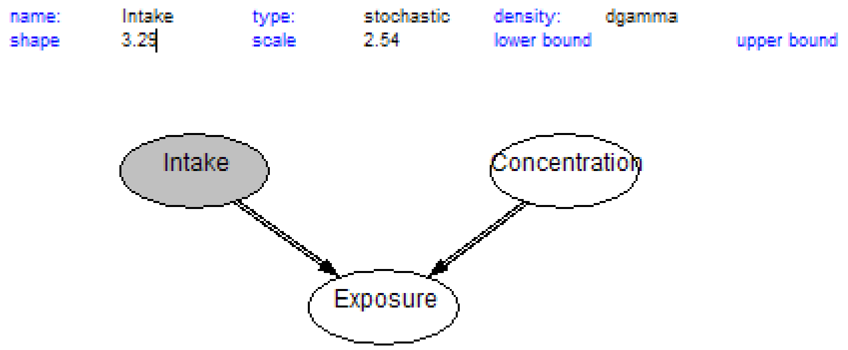


Figure 12. Simple OpenBUGS model of exposure, with details of the Intake variable displayed

Several steps using a combination of menus and dialog boxes were needed to prepare the model, which was run for 10,000 iterations, taking about 5 seconds. The summary statistics are shown in Table 5 and the density graph for exposure in Figure 12. There was no obvious way to obtain other statistics or a larger graph from the program. One statistic to note is the 'MC error'. This is an estimate of the error calculated by the program, which can be used to decide on the number of iterations needed. In this case it was small relative to the mean, so the number used was probably more than adequate.

Table 5. Summary statistics for simple exposure model in OpenBUGS

	Intake	Concentration	Exposure
Mean	1.298	2.018	2.615
Std dev	0.7206	2.008	3.287
MC error	0.007424	0.02284	0.03518
2.5th percentile	0.2973	0.04833	0.04516
5.0th percentile	0.3919	0.1024	0.09056
10.0th percentile	0.5052	0.2069	0.1998
25.0th percentile	0.7732	0.5957	0.5705
Median	1.164	1.424	1.516
75.0th percentile	1.685	2.789	3.375
90.0th percentile	2.274	4.648	6.311
95.0th percentile	2.658	6.077	8.869
97.5 percentile	3.046	7.377	11.65

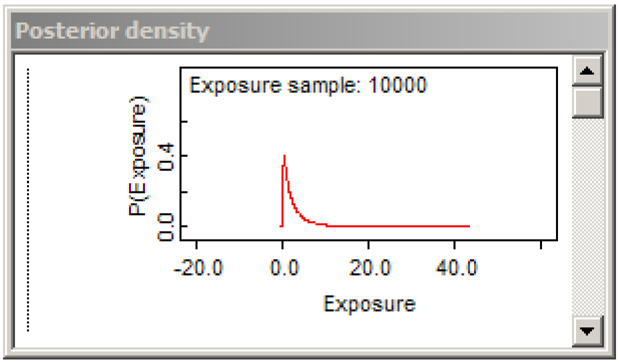


Figure 13. Density graph for exposure from simple model in OpenBUGS

Running the model was quite quick, and MCMC methods are very flexible with many features that were not required for the exposure simulation. However, the user interface of OpenBUGS is not easy to learn and use. For example, the dialog box from which the output options were selected had a drop-down list for the variables, but the names had to be typed into the list before they could be selected. It is clearly a powerful tool for a specialist, but not well-suited to this application.

4. References

Auer, M. (2012). MonteCarlito 1.10 --- Free Excel Tool for Monte Carlo Simulation (accessed 10/06/2011). http://www.montecarlito.com/

Crawley, M. J. (2007). The R book. John Wiley & Sons, Chichester, 942 pp.

Jensen, F. V. (1996). An introduction to Bayesian networks. UCL Press, London, UK, pp.

Palisade Corporation (2007). @RISK 5.0 risk analysis software. http://www.palisade.com/risk/

Varela, J. R. (2011). Simulación 4.0 (accessed 09/05/2011). http://www.ucema.edu.ar/u/jvarela/index_eng.htm

Probabilistic modelling for assessment of exposure via drinking water

Final Report of Project Defra WT1263 / DWI 70/2/273

Appendix B. Note on the correction of model likelihood estimates for transformed variables

The main report referred to the need to correct the maximum likelihood value from model fitting when the data were transformed. This short note gives a semi-formal outline of the method used.

Assume that X is a random variable (also known as an uncertain quantity) with probability density function f_X. Further assume that g is a transformation that is one-to-one and differentiable, and that $Y = g(X)$ with probability density function f_Y. It is easily shown that

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|}$$

where $y = f(x)$ and g' is the derivative of g (see Kendall *et al.*, 1987, 1.26 for a version of this result). If f_Y is defined by a parameter value α, the log likelihood of that value given the observations $y_1, \ldots, y_n$ (= $f(x_1),\ldots, f(x_n)$) is defined (Kendall & Stuart, 1979, 18.1) as

$$\log \mathcal{L}_Y(\alpha|y_1, \ldots, y_n) = \sum_1^n \log f_Y(y_i|\alpha)$$

It follows directly that

$$\log \mathcal{L}_X(\alpha|x) = \log \mathcal{L}_Y(\alpha|y) + \sum_1^n \log|g'(x_i)|$$

In the case of the square root transform, assuming the positive square root is taken,

$$g(x) = x^{1/2}$$
$$g'(x) = \frac{1}{2}x^{-1/2}$$

so

$$\sum_1^n \log|g'(x_i)| = -\frac{1}{2}\sum_1^n \log x_i - n\log 2$$

This result assumes that the data values are independent of the parameter value and known with certainty. This is not the case when using intervals (fitting to histograms) as above, or when the data set is 'censored' by the limit of detection as below The summation should be taken over the values that are specified without uncertainty, but additional correction terms are required for the censored or interval values (Kendall & Stuart, 1979, 32.17). Using constructed data sets with the R function *fitdestcens*, we found that no additional corrections were required in either case. In particular, when all of the values were gives as intervals, which was the case for the DWCS intake data, the likelihood given by *fitdestcens* was correct without adjustment.

References

Kendall, M.; Stuart, A. (1979). The advanced theory of statistics. Volume 2. Inference and relationship. Charles Griffin & Company, London. Fourth Edition, 748 pp.

Kendall, M.; Stuart, A.; Ord, J.K. (1987). Kendall's advanced theory of statistics. Volume 1. Distribution theory. Charles Griffin and Company, London. Fifth Edition, 604 pp.

Probabilistic modelling for assessment of exposure via drinking water

Final Report of Project Defra WT1263 / DWI 70/2/273

Appendix C. Distributions of contaminants in tap water samples

Of the 10 substances for which data from 2010 were available, the main report considered six. This appendix shows graphs of the distributions of all 10 substances (Table 1, Figure 1) for England and Wales as a whole, and illustrates the geographical variation in nitrate concentration (Figure 2).

The distributions are shown in the form of kernel density plots, which are, in effect, smoothed histograms that estimate the probability density function for the data represented by the sample. The smoothing function chosen is Gaussian (i.e. the normal distribution), so a single point would be artificially smoothed to the shape of a normal distribution. For the large data sets used here, smoothing should create few artefacts, except possibly at the extremes of the distribution.

Most of the sets contained some values recorded as less than the limit of detection (LoD), which varied between and within companies. In these graphs, the black curve shows the density derived from all the samples, with points recorded as <LoD shown at the LoD. The blue curve, where visible, shows the density derived after removing all the samples recorded as <LoD. The maximum of the concentration axis is the 99th percentile in each case of the set with values less than LoD removed.

Table 1. Determinands present in the data set for 2010.

Determinand	Prescribed concentration	Maximum permitted LoD	Units	Number of samples	Number < LoD
Sulphate*	250.0	25.0	mg	10,524	48
Sodium†	200.0	20.0	mg	18,234	7
Nitrate	50.0	5.0	mg	24,378	0
Iron†	200.0	20.0	µg	45,684	17,376
Copper†	2.0	0.2	mg	13,029	1,322
Arsenic	10.0	1.0	µg	12,825	3,958
Lead†	25.0	2.5	µg	12,667	6,323
Selenium†	10.0	1.0	µg	12,646	5,428
Chloride*	250.0	25.0	mg	10,536	2
Manganese†	50.0	5.0	µg	41,420	24,775

* Indicators † Considered in main report

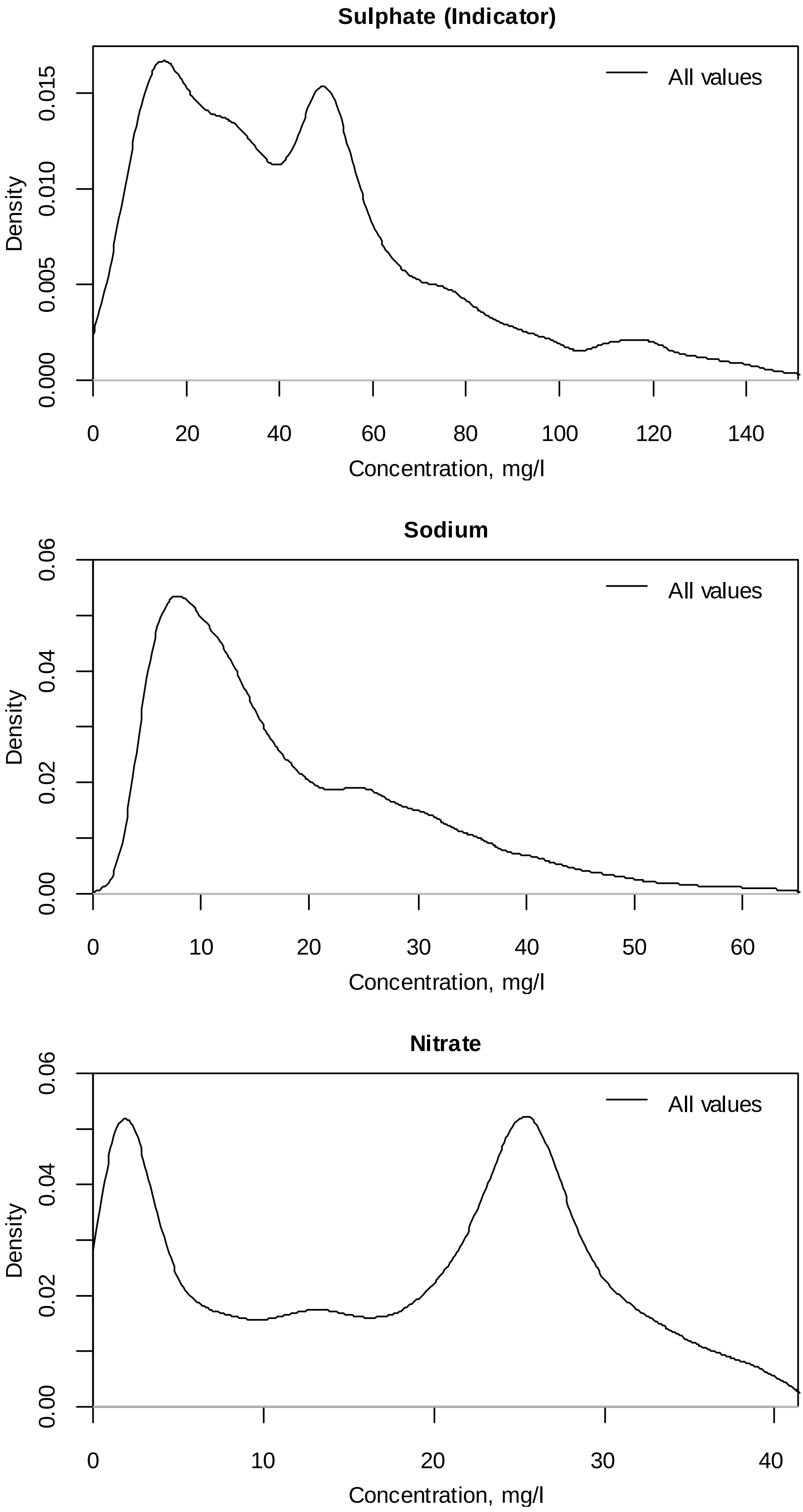
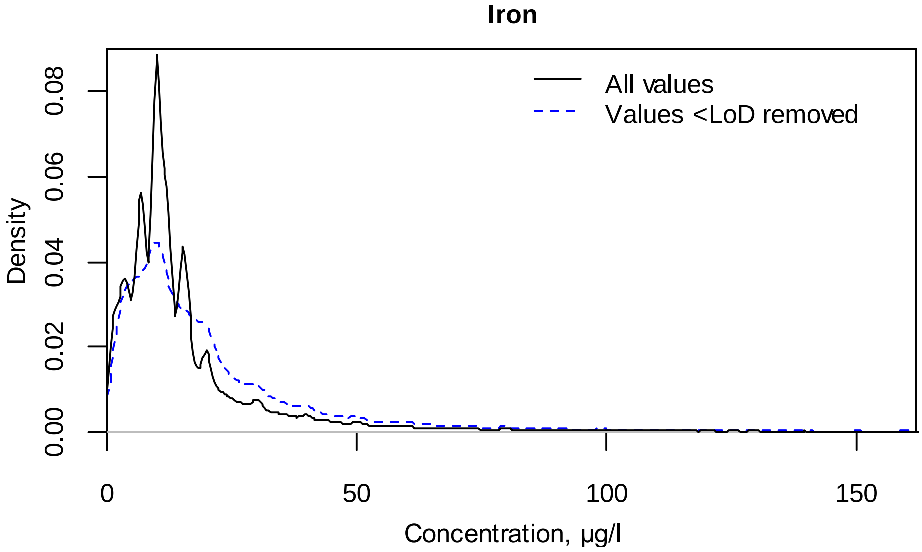
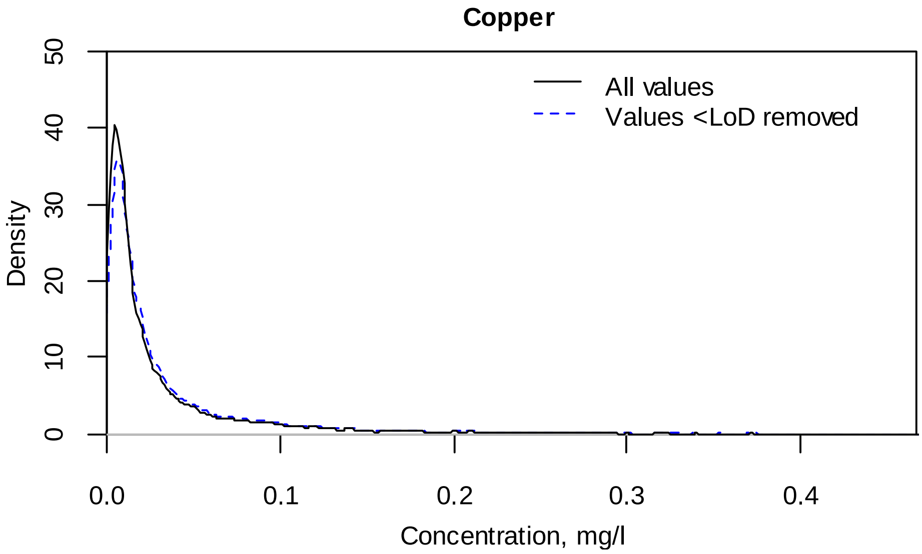


Figure 1. Estimated density functions for concentration in tap water in England and Wales, 2010





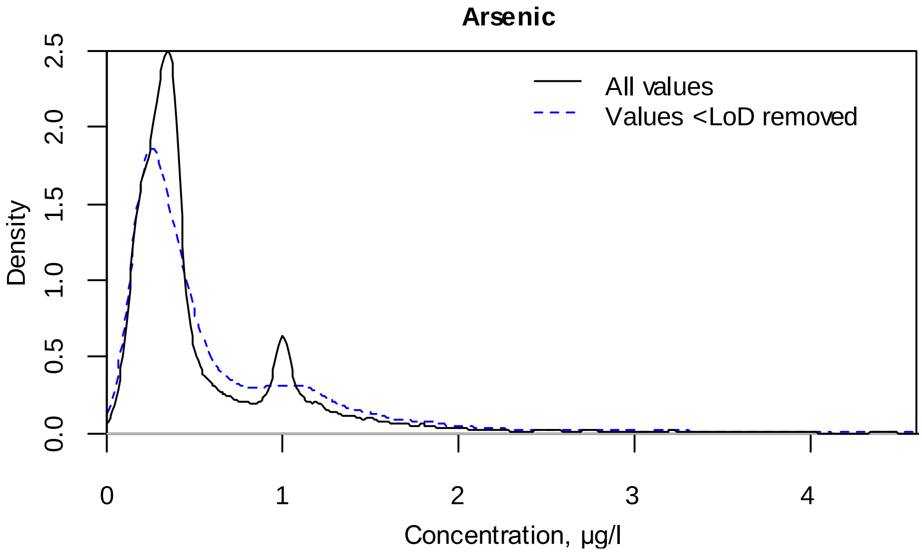


Figure 1 (continued). Estimated density functions for concentration in tap water in England and Wales, 2010

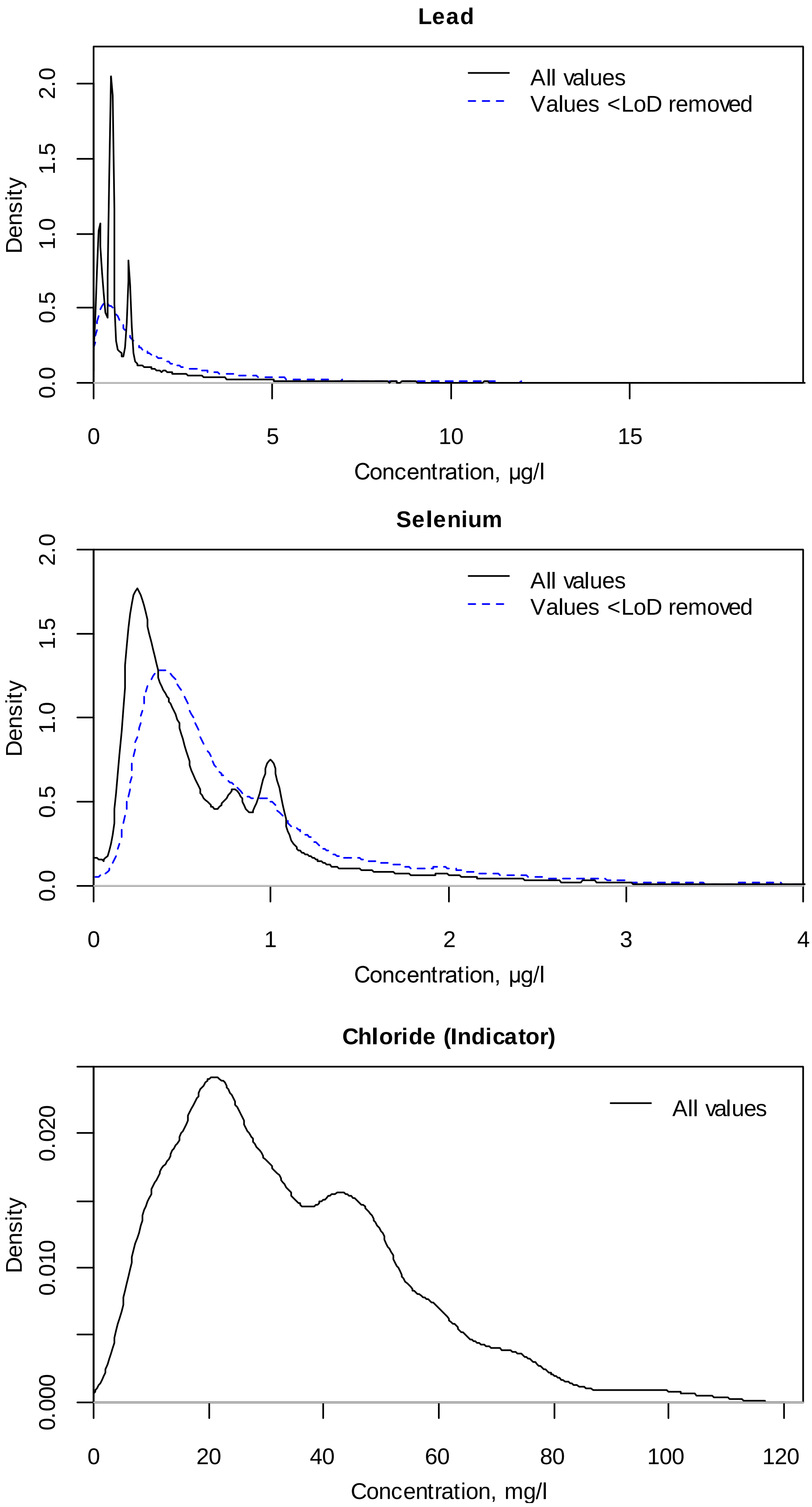


Figure 1 (continued). Estimated density functions for concentration in tap water in England and Wales, 2010

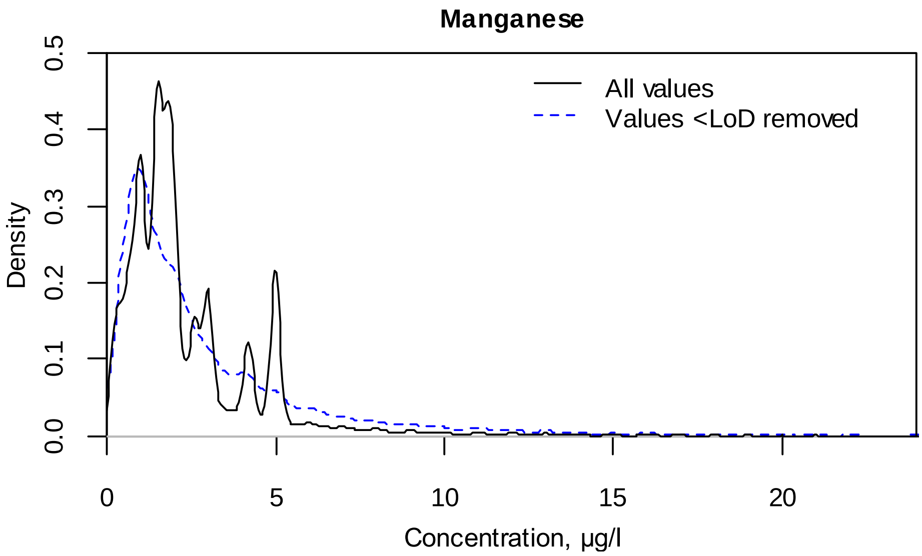


Figure 1 (continued). Estimated density functions for concentration in tap water in England and Wales, 2010

In contrast to the metals considered in the main report, sulphate and nitrate both have strongly bimodal distributions (distributions with two peaks). Chloride has a smaller second peak and there is some evidence of one for sodium also. The apparent multiple peaks for some of the other substances seem to be artefacts of multiple LoDs, though these could be masking patterns in the data. These provide clear evidence of multiple 'populations' in the data, probably due to geographical variations, such as topography, soil types and land use, and used of different sources, such as ground and surface water.

The regional variation in the distribution of nitrate concentration is illustrated below (Figure 2). The data from several companies were grouped into broad regions taking into account geographical proximity and similarity in the distributions. None of the regions used consisted of a single company.

It can be seen that the type of distribution varies widely. In Region A it resembles a lognormal distribution, with strong positive skew and a long tail. In Region G this is combined with a distribution with a mode around 10 mg/l, whereas in Region B there is a second distribution with a mode around 25 mg/l.

In the other regions, the mode tends to be in the range 25–30 mg/l, but only Region E has a clear single peak. All the other regions appear to contain several different concentration populations, possibly as many as four in Region D.

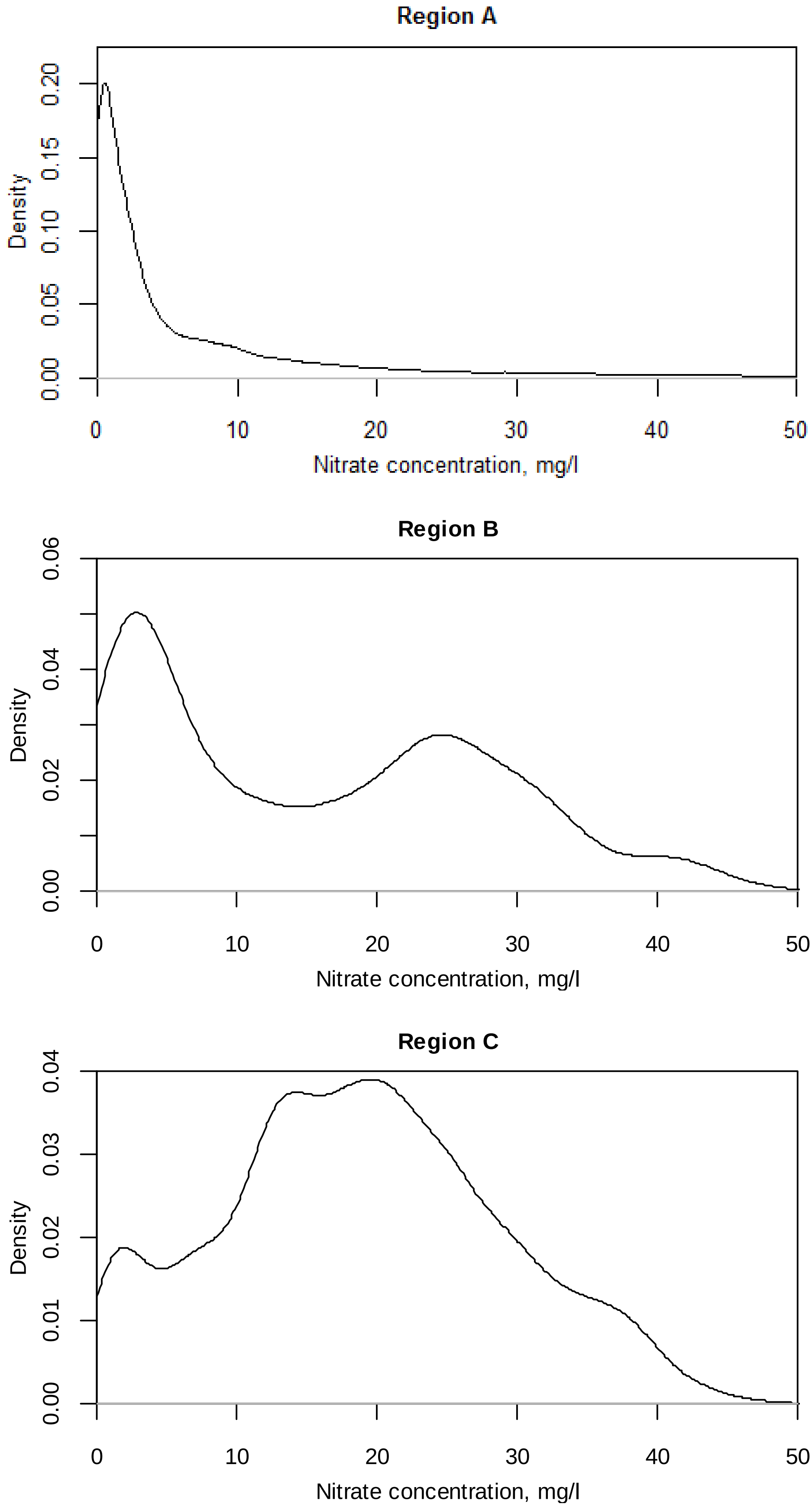


Figure 2. Estimated density function for nitrate concentration in tap water in regions of England and Wales, 2010

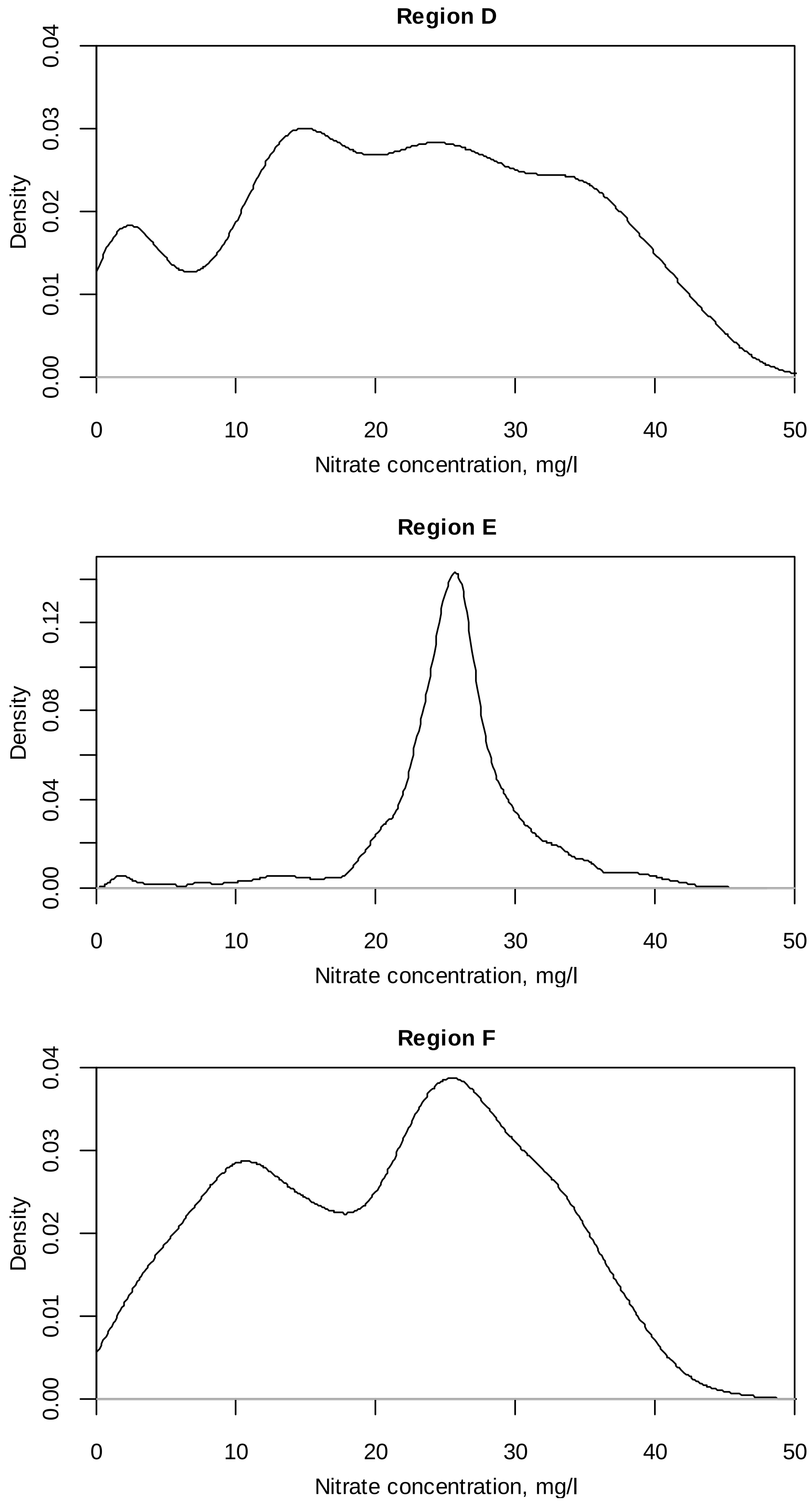


Figure 2 (continued) Estimated density function for nitrate concentration in tap water in regions of England and Wales, 2010

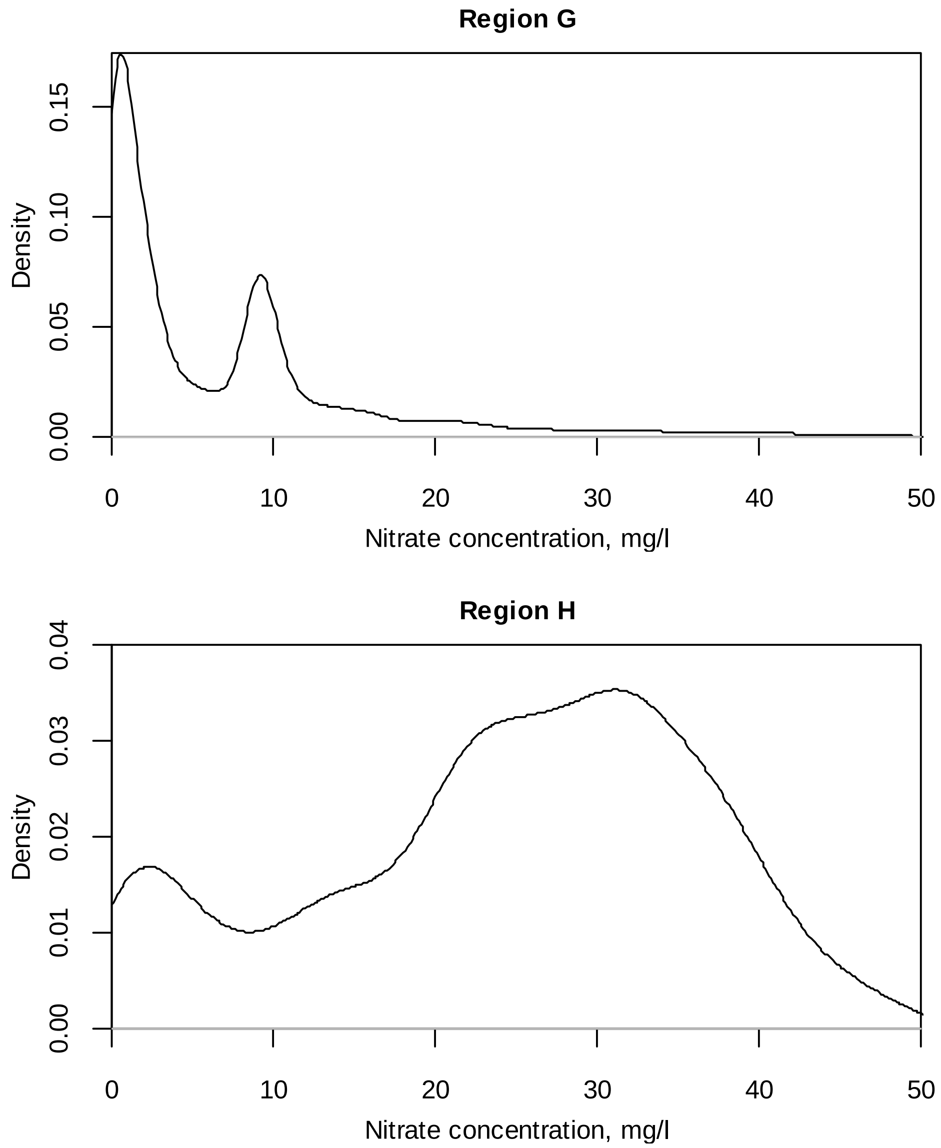


Figure 2 (continued) Estimated density function for nitrate concentration in tap water in regions of England and Wales, 2010